

BEST AVAILABLE COPY

Attorney's Docket No. 5800-2B

PATENT

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re: Glucksmann, et al.

Appl. No.: 09/464,685

Filed: 12/16/99

For: 2871 RECEPTOR, A NOVEL G-PROTEIN COUPLED RECEPTOR

Group Art Unit: 1635

Examiner: A. Wang



October 29, 2001

Commissioner for Patents
Washington, DC 20231

REPLY BRIEF

Sir:

This Reply Brief is filed in response to the Examiner's Answer, mailed 8/28/01.

Status of Amendments

An Amendment After Final Action was filed October 29, 2001 to cancel subject matter from claims 73 and 81. Applicants note that U.S. Patent No. 6,063,596 has issued from the PCT application cited by Applicants during prosecution (Int'l. Pub. No. WO 99/29849; see paper 8 including IDS initialed by Examiner in Office Action mailed 8/25/00). Because the claims of this issued patent read on the proposed claims of the present application, Applicants filed the amendment in order to reduce issues on appeal. A copy of the claims prior to this amendment can be found attached to the Appeal Brief filed in this case on July 10, 2001; the claims attached hereto are shown as if the changes suggested in the Amendment After Final Action had been entered.

Introduction

The Examiner has raised several new arguments in the Examiner's Answer and has renewed grounds for the rejection that were previously overcome by the Appellants. In the final Office Action, the Examiner states, "Applicants have not clearly demonstrated that the cloned nucleic acid and its encoded polypeptide is actually a GPCR as was noted in the utility rejection,"

but “Applicants do indeed provide multiple well established and specific utilities for a GPCR.” See, Office Action dated February 12, 2001, page 3. In the Examiner’s Answer, the Examiner agreed with Appellants’ statement of the issues in the Appeal Brief that the single issue was whether 2871 is a GPCR (Examiner’s Answer, page 2). However, the Examiner proceeds to set forth a new ground for the utility rejection, indicating that even if Applicants establish 2871 as a GPCR, members of the GPCR family of polypeptides do not have well-established utility. See, Examiner’s Answer, pages 4-5.

The Examiner also cites a new reference in the Examiner’s Answer in support of the argument that a protein’s sequence may not be used to predict its function (Attwood (2000) *Science* 290:417-473). The Examiner’s Answer further includes citations to two references that were cited in the first Office Action as a grounds for rejection under 35 U.S.C. § 101, but that were not cited in the final Office Action. The fact that these references were not cited in the final Office Action led Appellants to believe that the rejection had been overcome to the extent that it was based on these references, and therefore the Examiner’s arguments were not addressed in detail in the Appeal Brief.

Appellants note that this change in direction by the Examiner, which was not explained, is a practice which makes patent prosecution more difficult. This practice serves to obscure the basis for the rejection and runs the risk of unfairly prejudicing applicants’ nascent property rights in their patentable subject matter. As stated by the Federal Circuit in *In re Oetiker*, “[t]he examiner cannot sit mum, leaving the applicant to shoot arrows in the dark hoping to somehow hit a secret objection harbored by the examiner.” 977 F.2d 1443, 24 USPQ2d 1443, 1447 (Fed. Cir. 1992) (Plager, J., concurring).

Because the Examiner previously admitted that GPCRs have “multiple well-established and specific utilities,” Appellants did not fully address the utility of GPCRs in the Appeal Brief. It is requested that the rejection be withdrawn or prosecution be reopened to give Appellants a fair opportunity to respond to the new and renewed grounds of rejection. However, should the rejection not be withdrawn and prosecution not be reopened, Applicants here present these arguments in response to the Examiner’s new and revived grounds of rejection. Responses to the Examiner’s new and revived arguments are addressed below in section A, while the issue of

utility of the present invention is discussed below in section B.

Argument

2871 Encodes a G-Protein Coupled Receptor

A. The evidence presented by the Examiner does not address the methods used by the Appellants to determine 2871 receptor function.

1. Berendsen is directed to the de novo prediction of protein tertiary structure from primary structure, not the prediction of protein function based on the presence of conserved functional domains.

In the first Office Action and the Examiner's Answer, the Examiner cited Berendsen *et al.* (1998) *Science* 282:642-643 in support of the argument that protein activity predictions based on functional domains are unpredictable. Because this reference was not cited in the final Office Action, Appellants believed that the rejection had been overcome to the extent that it was based on this reference. However, the Examiner has cited Berendsen in the Examiner's Answer and thus it will be addressed here.

The teachings of Berendsen are directed to methods of predicting a protein's tertiary structure from its primary sequence. Berendsen states, "[t]he prediction of the native conformation of a protein of known amino acid sequence is one of the great open questions in molecular biology and one of the most demanding challenges in the new field of bioinformatics," and then proceeds to discuss computer simulations of protein folding. See Berendsen at 642. In the Examiner's Answer, the Examiner appears to acknowledge that Berendsen is not directed to the functional domain based predictions of protein function utilized by Appellants, but notes that "the activity of any protein or polypeptide is dependent on its structure." (Examiner's Answer, page 9) While Appellants agree that some regions of a protein must retain a certain conformation in order for the protein to be active, it does not follow that a protein's tertiary structure must be known in order to determine the activity of that protein. In fact, three-dimensional structures have been elucidated for only a very few of the thousands of proteins having known biochemical or physiological activity. Accordingly, the caveats regarding predictions of tertiary structure found in Berendsen are not relevant to methods for predicting

protein function used by the Appellants.

2. *Galperin is directed to context-based methods of predicting protein function, not to predictions of protein function based on the presence of functional domains.*

In the first office action, the Examiner cited Galperin *et al.* (2000), *Nature Biotechnology* 18:609-613, in support of the argument that a protein's function cannot be predicted from the presence of conserved functional domains. This reference was not cited in the final Office Action, leading Appellants to believe that the rejection was overcome to the extent that it was based on this reference. However, the Examiner has cited the reference in the Examiner's Answer and thus it will be addressed here.

The teachings of Galperin are directed to the prediction of protein function using comparative genomic approaches. The abstract for the Galperin reference states, "[s]everal recently developed computational approaches in comparative genomics go beyond sequence comparison. By analyzing phylogenetic profiles of protein families, domain fusions, gene adjacency in genomes, and expression patterns, these methods predict many functional interactions between proteins and help deduce specific functions for numerous proteins." The authors then proceed to discuss the strengths and weaknesses of these genomic context-based methods of functional prediction. Accordingly, the primary teachings of Galperin are not directed to the methods used by the Appellants to predict 2871 function.

In rebutting Appellants' arguments, the Examiner quotes Galperin: "sequence comparison methods, even the best ones, are of little help when a protein has no homologs in current databases or when all hits are to uncharacterized gene products." While Appellants agree that sequence similarity with uncharacterized gene products cannot be used to determine a protein's activity, this caveat does not apply to determine the function of the 2871 receptor. In the present case, the function of the 2871 receptor was determined based on the presence of sequence similarity with a conserved functional domain characteristic of the rhodopsin family of GPCR's. As described fully in Appellants' Appeal Brief and illustrated in Appendix E of the same, this conserved functional domain was elucidated from the sequences of a number of

rhodopsin-family GPCR's having known biochemical activities. Accordingly, Galperin's statement that "sequence comparison methods...are of little help...when all hits are to uncharacterized gene products" is true but does not undermine the reliability of the methods of functional prediction used by the appellants.

The only additional teachings that Galperin provides regarding prediction of protein function based on sequence similarity with proteins of known function are also supportive of the strength and reliability of the methods used in the present application. Galperin states (page 613, column 1) that comparative genomic methods of predicting protein function discussed in the reference "provide a useful extension of, and in a sense a genome-based framework for, sequence and structural methods which remain the cornerstone of computational genomics." Thus, Galperin distinguishes between the comparative genomics-based methods of functional prediction reviewed in the reference and the pattern based methods for functional prediction used by the Appellants, and further demonstrates that the authors consider the approach used by Appellants to be reliable.

3. *Attwood distinguishes between the reliability of module-based prediction of protein function and pattern-based prediction of protein function and presents arguments supporting the diagnostic reliability of pattern databases.*

The Examiner's Answer includes a citation to a new reference (Attwood (2000) *Science* 290:471-473) in support of the argument that sequence similarity cannot be used to predict protein function. Specifically, the Examiner cites the statement: "[i]f the best hit in a database search is a match to a single domain module, it is unlikely that the function annotation can be propagated from the parent protein to the query sequence," and "[t]he presence of a module tells little of the function of the complete system; knowing most of the components of a mosaic does not allow us easily to predict a missing one, and modules in different proteins do not always perform the same function." Attwood (2000) *Science* 290 at page 472, column 2. A careful reading of the Attwood reference makes it clear that these statements refer not to the prediction of protein function based on the presence of a conserved functional domain, but rather to the prediction of function based on the presence of a single motif or module. Such modules are

defined by Attwood as "autonomous folding units that often function as protein building blocks, forming multiple combinations of the same module or mosaics of different modules." *Id.* In the present case, Appellants have determined the function of the 2871 receptor based on the fact that 255 contiguous amino acids of the 2871 polypeptide provide an excellent fit to the empirically-derived model of the GPCR family that includes rhodopsin. This statistical model is not solely based on the presence of a single autonomous folding unit.

The differences between the reliability of motif or module-based methods of protein function prediction and functional domain-based methods of function prediction are discussed in greater detail in Attwood (2000) *Int. J. Biochem. Cell Biol.* 32:139-155 ("*IJBCB*," provided as Appendix G), a more comprehensive review article published by Attwood in the same year as the reference cited by the examiner. In this reference, Attwood teaches that while functional prediction methods based on the presence of a single motif may be problematic because matches to single motifs lack biological context (see Attwood, *IJBCB* at 144), many of the flaws inherent in these single motif-based methods are overcome in pattern databases such as Pfam. Attwood states:

"[p]attern databases offer several benefits: (i) by distilling multiple sequence information into family descriptors, trivial errors in the underlying sequences may be diluted; (ii) annotation errors may be quickly spotted if the description of one sequence differs from that of its family; and (iii) they allow specific diagnoses, placing individual sequences in a family context for a more informed assessment of possible function."

Attwood, *IJBCB* at 153.

Attwood also teaches the diagnostic advantages of manually-generated databases such as Pfam (which is based on hand-edited seed alignments; see Attwood, *IJBCB* at 149). Attwood states, "manually annotated databases are set apart from their automatically created counterparts by virtue of (i) providing *validation* of results and (ii) offering detailed information that helps to place conserved sequence information in structural or functional contexts." (Attwood, *IJBCB* at 152). Attwood further states that while pattern databases are small in comparison with sequence repositories, "their diagnostic potency ensures that pattern databases will pay an increasingly

important role as the post-genome quest to assign functional information to raw sequence data gains pace." (Attwood, *IJBCB* at pp. 153-154) Thus the teachings by Attwood regarding pattern databases, particularly manually-generated pattern databases, are strongly supportive of the reliability of these techniques.

Thus, the Examiner seizes on a single brief review article by Attwood about caveats of sequence comparison methods to discredit sequence comparison methods in general (Examiner's Answer, page 7, "protein function cannot be ascertained from analysis of its components.") Applicants agree generally with Attwood's argument in the new reference cited by the Examiner that predictions of protein function based on a single motif are not necessarily reliable. However, those of skill in the art distinguish between the presence of a single motif in a protein and the presence of configurations of multiple motifs, or a pattern, which is diagnostic of a particular protein family.

Attwood has published a number of articles describing patterns that are diagnostic of G-protein coupled receptors¹, and is known as one of the creators of the PRINTS sequence comparison method and database. Perhaps most pertinent here is an article published by Attwood after the article cited by the Examiner, entitled: "A compendium of specific motifs for diagnosing GPCR subtypes." Attwood (2001) *TRENDS in Pharmacological Sciences* 22(4): 162-165 ("*TiPS*," provided as Appendix H). In this article, Attwood discusses the differences between several sequence comparison methods and describes the use of her PRINTS methods and database for the analysis of GPCRs (available at <http://bioinf.man.ac.uk/cgi-bin/dbbrowser/fingerPRINTScan/muppet/FPScan.cgi>, as indicated in Figure 1). See Attwood, *TiPS* at 164.

A PRINTS analysis of the closest publicly disclosed polypeptide sequence to the subject of the present application (*i.e.*, the sequence disclosed in U.S. Patent No. 6,063,596 as SEQ ID NO: 3) shows an identification of the "GPCRRHODOPSN" fingerprint, with an E-value of 3.1e

¹ Attwood's work includes: Attwood and Beck (1994), *Protein Eng.* 7(7): 841-848, entitled "PRINTS—a protein motif fingerprint database"; Attwood and Findlay (1994) *Protein Eng.* 7(2): 195-203, entitled "Fingerprinting G-protein Coupled Receptors"; Attwood *et al.* (1991) *Gene* 98(2): 153-159, entitled "Multiple Sequence Alignment of Protein Families Showing Low Sequence Homology: A Methodological Approach Using Database Pattern-matching Discriminators for G-protein-linked Receptors."

²⁹ and a P-value of $1.2e^{-34}$ (see output, attached as Appendix I). As indicated in the documentation for PRINTS also available at this site, “[t]he reported P-value of any fingerprint result is the product of the p-values for each motif. The motif p-values represent the probability that a comparison between the motif and a random sequence would achieve a score greater than or equal to the score attributed to the match between your query sequence and the motif.” The E-value is the expected number of occurrences of sequences scoring greater than or equal to the query’s score. Thus, the very low P-value and E-value obtained from Attwood’s PRINTS analysis concurs with the Pfam diagnosis described by Applicants that the 2871 sequence is a GPCR. Accordingly, the Examiner’s use of Attwood to discredit sequence comparison methods in general is inconsistent with Attwood’s work, which strongly supports the conclusion that the 2871 sequence is a GPCR.

4. *The Examiner’s failure to credit the predictive power of sequence comparison methods is at odds with accepted practice in the art.*

The Examiner notes (Examiner’s Answer, paper number 18 mailed 8/28/01, page 3) that “[m]oreover, the specification discloses that the cloned GPCR shares a high score with the seven transmembrane rhodopsin family,” and further states on page 4 that “the specification notes that proteins with putative seven transmembrane domains, much like applicants, are not necessarily GPCRs such as *boss* and *fz* cloned from *Drosophila*.” The Examiner also states (Examiner’s Answer, page 6-7) that “Figure 2 provides for only the DRY triplet and low sequence homology.” Based partly on this line of reasoning, the Examiner asserts that the specification lacks “a specific and substantial utility [and] a well established utility.”

This line of reasoning by the Examiner is inconsistent with the understanding of one of skill in the art of Pfam alignments, and of sequence comparisons in general. As known to those of skill in the art (and described in the Pfam documentation available at <http://pfam.wustl.edu/faq.shtml>), Pfam alignments do not display homology between pairs of sequences but rather display the fit of a particular query sequence to a particular protein family model. As discussed on the Pfam “Help Page:FAQ” available at the address above, complaints [like the Examiner’s present complaint] about the quality of the alignments generally arise

“because people aren’t used to looking at multiple alignments of hundreds or thousands of sequences. Remember that a rare insertion in even just one sequence [in the protein family] means having to open a gap in the whole alignment: Pfam full alignments look very gappy for this reason, but in fact they’re not.”

The Examiner also ignores that *boss* (bride of sevenless) and *fz* (frizzled) show low similarities to GPCR domains in Pfam alignments. One of skill in the art understands that Pfam alignments of *boss* and *frizzled* with the highest-scoring seven transmembrane domain models for each (7tm_3 and 7tm_2, respectively) have negative scores. In contrast, the 2871 sequence has a high positive score for the rhodopsin subfamily that is described by Pfam model 7tm_1. Pfam “bit scores” represent the log base 2 of a ratio. In the numerator of this ratio is the probability of the sequence given the hypothesis that the sequence belongs to the protein family being modeled. In the denominator of this ratio is the probability of the sequence given the hypothesis that the sequence was generated according to a random background model. Thus, the bit score of 183 for protein 2871 with the Pfam 7tm_1 model means this protein sequence is 2^{183} times more likely to be observed if it were generated by the 7tm_1 model than if the sequence were generated by the other model. We note that 2^{183} (about 1.2×10^{55}) greatly exceeds the estimated number of atoms comprised by the planet Earth. In contrast, the optimal score for *boss* to a GPCR family is -53, and the optimal score for *frizzled* is even lower, at -112. In other words, the sequence of *boss* is 2^{53} times more likely to be observed if it were generated by the random background model than if it were generated by the best-fitting GPCR model. Although 2^{53} does not exceed the estimated number of atoms that are comprised by the planet Earth, we note that 2^{53} is an extremely large number (about 9×10^{15}). Thus, contrary to the Examiner’s arguments, the fact that the *boss* and *frizzled* proteins have seven transmembrane domains does not detract from Applicants’ evidence that the sequences of the present invention are GPCRs.

The Examiner has attacked Applicants’ use of sequence comparison methods by quoting caveats largely out of context. As one of skill in the art is aware, any methodology is fallible to some degree and there are always exceptions to a rule; thus, most if not all articles describing sequence comparison methods also discuss the shortcomings of those methods. The Examiner seizes on these caveats to discredit the use of sequence comparison methods. The Examiner’s

approach is at odds with that of the art, which has embraced sequence comparison methods, particularly as those methods have advanced in sophistication with the rapid advances of the genomic era.

A brief survey of PubMed (accessible at <http://www.ncbi.nlm.nih.gov/>) shows dozens of peer-reviewed, scientific articles published every month describing novel discoveries of sequences having strong identity to sequences of known function. The acceptance of sequence comparison methods by the art is evidenced in many places. For example, Mount (2001) *Bioinformatics: Sequence and Genome Analysis* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York), page 282 (provided as Appendix J) states that “[d]atabase similarity searches have become a mainstay of bioinformatics.” Mount goes on to explain that, “[a]s a rough rule, if more than one-half of the amino acid sequence of query and database proteins is identical in the sequence alignments, the prediction is very strong. As the degree of similarity decreases, confidence in the prediction also decreases. The programs used for these database searches provide statistical evaluations that serve as a guide for evaluation of the alignment scores.” As noted by Gusfield (1997) *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology* (Cambridge University Press, New York, New York), at pages 212-213 (provided as Appendix K),

[s]equence comparison, particularly when combined with the systematic collection, curation, and search of databases containing biomolecular sequences, has become essential in modern molecular biology. * * * The first fact of biological sequence analysis: In biomolecular sequences (DNA, RNA, or amino acid sequences), high sequence similarity usually implies significant functional or structural similarity. Evolution reuses, builds on, duplicates, and modifies “successful” structures (proteins, exons, DNA regulatory sequences, morphological features, enzymatic pathways, *etc.*).

* * *

‘Today, the most powerful method for inferring the biological function of a gene (or the protein that it encodes) is by sequence similarity searching on protein and DNA sequence databases. With the development of rapid methods for sequence comparison, both with heuristic algorithms and powerful parallel computers, discoveries based solely on sequence homology have become routine.’ [citation omitted] * * * It is now standard practice, whenever a new gene is

cloned and sequenced, to translate its DNA sequence into an amino acid sequence and then search for similarities between it and members of the protein databases.”

Another indicator of the importance of sequence comparison methods to the “new paradigm” of modern molecular biology is the fact that the most-cited paper of 1990-1998 is the publication describing BLAST: Altschul (1990) *J. Mol. Biol.* 215: 403, entitled “Basic Local Alignment Search Tool.” (citation figures available at <http://www.isinet.com/isi/hot/research>) (provided as Appendix L). Accordingly, the Examiner’s efforts to discredit sequence comparison methods in general is inconsistent with the art, which supports the use of sequence comparison methods and thus the conclusion that 2871 is a GPCR.

5. *Despite the fact that the histamine receptor family is divergent, members of these families were identified as GPCRs based on sequence similarity with known GPCRs.*

In the Appeal Brief, Appellants cited Nguyen *et al.* (2001) *Mol. Pharmacol.* 59:427-433 which describes the identification of the histamine receptor H4 based on sequence similarity with known GPCRs. In response, the Examiner has cited the teaching by Nguyen that the histamine receptors H₁, H₂, and H₃ share less than 35% identity with one another and each has greater identity with other aminergic receptors. This statement by Nguyen supports rather than discredits the reliability of the methods of functional prediction used by the Appellants, as it demonstrates that the activity (in this case the G-protein mediated signal transduction activity) of a protein can be predicted based on sequence identity of less than 35%.

Despite the fact that histamine receptors share only moderate sequence identity with each other, the H₁, H₂, H₃, and H₄ receptors were each recognized as being a G-protein coupled receptor having G-protein mediated signal transduction activity based on sequence identity. For example, Yamashita *et al.* (1991) *Biochem.* 88:11515-11519 (provided as Appendix M) describe the cloning of the H₁ receptor and note that “[t]he histamine H₁ receptor is highly similar to other G protein-coupled receptors.” Yamashita at 11518. Similarly, Gantz *et al.* (1991) *Proc. Natl. Acad. Sci.* 88:429-433 (provided as Appendix N) describe the cloning of the H₂ receptor and

note that "comparison of the deduced amino acid sequence to that of other G-protein-linked receptors with presumed seven-transmembrane motifs revealed extensive homology." The H₃ histamine receptor was identified and cloned based on a high degree of sequence similarity with biogenic amine GPCRs (Lovenberg *et al.* (1999) *Mol. Pharmacol.* 55:1101-1107, provided as Appendix O). Finally, as described in the Appeal Brief, Nguyen *et al.* describe the cloning of the H₄ receptor based on a query of GenBank to identify sequences sharing sequence similarity with GPCRs. Thus, the G-protein mediated signal transduction activity of all of the histamine receptors was accurately predicted based on sequence similarity with known GPCRs. Accordingly, the Examiner's attempt to use Nguyen to discredit functional prediction methods is misguided.

6. *The tumor suppressor activity of p73 was predicted based on sequence identity with the known tumor suppressor p53.*

In the Appeal Brief, Appellants cite Dickman *et al.* (1997) *Science* 277:1605-1606, which teaches that the tumor suppressor activity of the p73 polypeptide was determined based on sequence similarity with the transcription activation, DNA-binding, and oligomerization domains of the known tumor suppressor protein p53. In response, the Examiner argues that Dickman teaches that the p73 gene is deleted in certain cancers. However, a careful reading of Dickman finds that the original determination of p73 protein's tumor suppression activity was made on the basis of sequence similarity alone. Dickman teaches that p73 was identified in a screen for genes that respond to certain immune system regulators. Dickman states, "[w]hen the French team sequenced the many potential targets their screen had turned up, they were shocked to find out that one false positive had remarkable similarities to p53." Dickman at 1605. It was only after p73's tumor suppression activity had been predicted on the basis of sequence similarity with p53 that the investigators thought to look for alterations in the p73 gene in cancer patients. Thus, Dickman is an excellent example of the value of sequence comparison-based approaches to discovery of new genes.

7. *Kliwer demonstrates the successful identification of novel nuclear receptors based on sequence similarity with the ligand-binding domain of known nuclear receptors.*

In the Appeal Brief, Appellants cite Kliwer *et al.* (1998) *Cell* 92(1): 73-82 as an additional example of the accurate determination of a protein's function based on the presence of functional domains. The Examiner seeks to discredit this argument by noting that the PXR.1 amino acid sequence is identical to the PXR.2 amino acid sequence except for a 41 amino acid deletion resulting from alternative splicing. This statement misses the point of the reference, which does not teach the isolation of the PXR.2 coding sequence based on the PXR.1 coding sequence but instead describes the cloning of both the PXR.1 and PXR.2 coding sequences based on sequence identity with motifs characteristic of known nuclear receptors. Kliwer states, "[i]n an effort to identify new member of the nuclear receptor family, we performed a series of motif searches of public EST databases. These searches revealed a clone . . . that had homology to the ligand-binding domain of a number of nuclear receptors." Kliwer at 74. Kliwer teaches that this EST was then used to clone the nuclear receptor PXR.1 and its splice variant PXR.2. Thus, Kliwer describes yet another successful use of sequence similarity with functional domains to predict protein function.

B. The evidence presented by Applicants supports a finding that the present invention satisfies the requirement of utility.

Applicants again note that these arguments are presented for the first time on appeal because the Examiner earlier indicated that the only issue was whether the disclosed sequence actually was a GPCR. Now, the Examiner asserts that even if the disclosed sequences are GPCRs, utility is not established. Because the Examiner has changed the utility rejection, Applicants have not had the opportunity to fully address the Examiner's arguments. Applicants here present these arguments in response to the Examiner's new and revived grounds of rejection.

1. *The 2871 receptor is useful in selectivity screening and therefore has a "well-established" utility.*

The Examiner has rejected claims 73, 74, 81, and 88-96 under 35 U.S.C. §101 on the grounds that the claimed invention "lacks patentable utility." (Feb. 12, 2001 Office action page 3). This does not correctly reflect the view in the art, where it is known that "[h]istorically, the superfamily of GPCRs has proven to be among the most successful drug targets and consequently these newly isolated orphan receptors have great potential for pioneer drug discovery." Stadel *et al.* (1997) *Trends Pharmacol. Sci.* 18:430-436; provided as Appendix P). Those of skill in the art recognize that the identification of a novel member of the G-protein coupled receptor family provides an immediate benefit. In addition to serving as reagents and targets in the diagnosis and treatment of 2871-mediated disorders as described in the specification on page 48 *et seq.*, all members of the GPCR protein family have utility in selectivity screening of candidate drugs that target GPCRs. It is known in the art that the clinical usefulness of a therapeutic compound is determined not only by its ability to bind and modulate a molecular target of interest, but also by its selectivity. Drugs that bind selectively to their molecular target are highly preferred over those that bind to structurally-related molecules, as the selective compounds are far less likely to have unwanted side effects in clinical use. See, for example, Hartig (1993) *NIDA Res. Monogr.* 134: 58-65, entitled, "The use of cloned human receptors for drug design," provided as Appendix Q; Fraser (1995) *J. Nucl. Med.* 36 (6 Suppl): 17S-21S, provided as Appendix R. Thus, an important component of any drug development strategy is determining the selectivity of the candidate drug for the molecular target of interest over structurally-related polypeptides. The effectiveness of selectivity screening in uncovering interactions that may result in undesirable clinical side-effects increases in proportion with the number of structurally-related polypeptides screened. In this situation, the usefulness of these structurally-related polypeptides is not dependent on their biological role or ligand-binding properties; their utility comes from the fact that they share significant sequence identity with the molecular target of the candidate drug.

An example of the use of orphan receptors in selectivity screening is found in Goodwin *et*

al. (2000) *Molecular Cell* 6:517-526, provided as Appendix S. This reference is directed to the identification of a specific agonist for FXR, an orphan nuclear receptor that regulates bile acid synthesis and is a target in the treatment of cholestasis. (See generally, Niesor *et al.* (2001) *Curr. Pharm. Des.* 7: 231-259). Goodwin states that many previously-identified FXR ligands interact with other proteins including bile-acid-binding proteins and transporters (Goodwin at page 518, column 1). In order to identify a compound that selectively modulates FXR, the authors screened for compounds that modulated FXR activity and then tested these compounds for their ability to activate other nuclear receptors that share structural similarity with FXR. Figure 1C of Goodwin shows that the compound GW4064 potently activates FXR but does not modulate the activity of the other nuclear receptors tested. Note that the nuclear receptor panel screened in Figure 1C includes the orphan nuclear receptors SHP-1 and LRH-1 in addition to receptors having previously-identified ligands, illustrating that studies often include orphan receptors.

More than 50% of prescription drugs act at GPCR targets, further showing the importance of GPCRs in screens for effective drugs. However, some of these drugs have efficacy problems and limiting side-effects because the compounds do not differentiate between receptor subtypes. See generally, Stadel *et al.*, (1997) *Trends Pharmacol. Sci.* 18: 430 (Appendix P); Lee and Kerlavage (1993) *Molecular Biology of G-Protein-Coupled Receptors*, 6 DN&P 488 (provided as Appendix T). Accordingly, because the GPCR protein family includes a number of key drug targets, members of this family share a common use in the selectivity screening of candidate drugs. The 2871 receptor shares a high degree of identity with the rhodopsin family of GPCRs (see specification Figure 2). This rhodopsin GPCR family includes targets for the treatment of numerous disorders including depression, anxiety, migraine, asthma, hypertension, and cardiovascular disorders. Thus, all members of this important class of GPCRs, including those disclosed in the present invention, have a specific, immediately available, real world utility in the selectivity screening of drugs directed at GPCR targets.

The 2871 receptor shares a high degree of identity with the rhodopsin family of GPCRs and is expressed in tissues including those of particular clinical significance to hematological disorders, such as hematopoietic cells (see Figure 7; see also Figures 5-6 and specification pages 6 and 19). Indeed, the 2871 gene is expressed at significant levels in all blood cell progenitors

analyzed by the inventors. It is highly expressed in bone marrow (CD34⁺), G-CSF-mobilized peripheral blood (containing circulating progenitors derived from bone marrow) and is moderately expressed in CD34⁺ adult bone marrow and CD34⁺ cord blood cells. It is also highly expressed in megakaryocytes as well as CD41⁺ (CD14⁺) bone marrow cells. G-CSF-mobilized peripheral blood contains circulating progenitors derived from bone marrow. Accordingly, expression of the 2871 gene is relevant for treating disorders associated with the formation of differentiated and/or mature blood cells, such as anemia, neutropenia, and thrombocytopenia.

The therapeutic and economic benefits that can result from selectivity screening are well known. One example is the events of 1994-1997 leading to Merck's marketing of the painkiller Vioxx, described in Gardiner Harris, *The Cure: With Big Drugs Dying, Merck Didn't Merge—It Found New Ones*, The Wall Street Journal, January 10, 2001, at A1 (provided as Appendix U). Merck's search for a novel pharmacologically suitable painkiller made use of *in vitro* screens to find drugs that inhibited the activity of Cox-2 but not Cox-1. Such drugs would inhibit prostaglandin production in most of the body but not the gut, thereby ameliorating pain while avoiding undesirable side effects. Candidate drugs from a collection of hundreds of synthesized drugs were first subjected to *in vitro* screening; a much smaller number of successful *in vitro* candidates advanced to *in vivo* screening in mice, and two successful nontoxic drugs from the mouse *in vivo* screens were advanced to even more expensive human clinical trials. Only one of these two drugs showed efficacy in clinical trials, ultimately received FDA approval, and is now being marketed as Vioxx. This example illustrates how a "real world" benefit can be obtained from distinguishing gene family members.

2. *The 2871 sequence has a high degree of identity to other sequences which have utility; therefore, the 2871 sequence has utility.*

The USPTO utility examination guidelines state, "[w]hen a class of proteins is defined such that the members share a specific, substantial, and credible utility, the reasonable assignment of a new protein to the class of sufficiently conserved proteins would impute the same specific, substantial, and credible utility to the assigned protein." 66 Fed. Reg. 1096. In

the present application, Applicants have demonstrated that the 2871 receptor is a member of the rhodopsin family of G-protein coupled receptors. Members of this family of receptors are known by those of skill in the art to share a specific, substantial, and credible utility. In fact, it has come to our attention that a U.S. patent has issued from an international application disclosed by Applicant in the Supplemental IDS returned by the Examiner with paper 8 (the Office Action mailed 8/25/00). In U.S. Patent No. 6,063,596, (the '596 patent) with inventors Lal *et al.* and assigned to Incyte Pharmaceuticals, issued 16 May 2000, one of the disclosed sequences has 98% identity to Applicant's 2871 sequence. The claimed invention of the '596 patent is described as providing human G-protein coupled receptors associated with immune response. Applicants' present claims are directed to methods of using the 2871 sequence of the present invention. Because there is an issued U.S. patent with claims to sequences with a high degree of identity to Applicant's 2871 sequences, the Patent Office must have found these sequences to have utility. Accordingly, a rejection of Applicants' present claims for lack of utility is inappropriate and should be withdrawn.

3. *The present invention is useful in its currently available form.*

The Examiner has stated that the specification does not provide "any evidence or guidance suggesting the claimed protein's activity" (Examiner's Answer at page 3) and that therefore doubt is cast on "whether the nucleotide sequence or its encoded protein can be used in **any** of applicants asserted utilities." (emphasis added; Examiner's Answer at page 4). Applicants disagree. As discussed in the specification and known in the art, GPCRs (G-protein coupled receptors) are responsible for G-protein mediated signal transduction. "GPCRs, along with G proteins and...intracellular enzymes and channels modulated by G-proteins, are the components of a modular signaling system that connects the state of intracellular second messengers to extracellular inputs." (specification page 2; see also pp. 6, 7, 20).

While the Examiner's assertion of lack of utility may reflect the thinking of the pre-genomics era, it does not accurately describe the current state of the art in drug discovery. Those of skill in the art appreciate that rapid advances in technology have led to dramatic changes in the

way research is conducted in many biomedical-related areas. “Molecular biology has had a dramatic influence” on active drug discovery and research projects in the pharmaceutical industry, particularly those involving GPCRs. See Stadel *et al.* (1997) *Trends Pharmacol. Sci.* 18:430-436; provided as Appendix P). The advances in molecular biology have led to what those in the art consider a “paradigm shift” in the way research and drug discovery is conducted. *Id.* In this new paradigm, the starting point in the process is the identification of new members of gene families such as the GPCR superfamily by “computational or bioinformatic methodologies.” Stadel at 430. “Once new members of the GPCR superfamily are identified, the recombinantly expressed receptors are used in functional assays to search for the associated novel ligands. The receptor-ligand pair are then used for compound bank screening to identify a lead compound that, together with the activating ligand, is used for biological and pathophysiological studies to determine the function and potential therapeutic value of a receptor antagonist (or agonist) in ameliorating a disease process.” Stadel at 434; see also Fraser (1995), *J. Nucl. Med.* 36 (6 Suppl): 17S-21S (Appendix R). Often, these screens are implemented in high-throughput format. *See id.* Thus, in the molecular biology field of the present invention, the discovery of a novel sequence is the key step, or “first link” of *Cross*. *See, Cross v. Iizuka*, 753 F.2d 1040, 1051 (Fed. Cir. 1985) (holding that “[w]e perceive no insurmountable difficulty, under appropriate circumstances, in finding that the first link in the screening chain, *in vitro* testing, may establish a practical utility for the compound in question.”)

Similarly, in drug development, the key step or “first link” is the discovery of a novel sequence such as that of the present invention; subsequent screening steps are routinely performed. As those in the art note, “the potential reward of using this [“reverse molecular pharmacological strategy”] approach is that resultant drugs naturally will be pioneer or innovative discoveries, and a significant proportion of these unique drugs may be useful to treat diseases for which existing therapies are lacking or insufficient.” Stadel at 434.

As those in the art are aware, much is known about GPCRs but many details of GPCR activity remain to be resolved, including comprehensive information about the mechanisms and domains of previously discovered GPCRs. Despite this lack of encyclopedic knowledge about GPCRs, members of this gene family have been shown to bind a variety of ligands and have been

successfully used for drug discovery. See, for example, Stadel *et al.*, (1997) *Trends Pharmacol. Sci.* 18:430. “Because of the proven link of GPCRs to a wide variety of diseases and the historical success of drugs that target GPCRs, we believe that these orphan receptors are among the best targets of the genomic era to advance into the drug discovery process.” Stadel at 436. “The fact that GPCRs mediate a broad spectrum of cellular events make these proteins an ideal target for drug interaction and therapeutics.” Lee and Kerlavage (1993) *Molecular Biology of G-Protein-Coupled Receptors*, 6 DN&P 488 (Appendix T).

4. *The rejection of the claims under 35 U.S.C. §101 and §112, first paragraph, is inconsistent with USPTO guidelines and supporting case law.*

The Utility Examination Guidelines state, “Applicant[s] need only provide one credible assertion of specific and substantial utility for each claimed invention to satisfy the utility requirement.” 66 Fed. Reg. 1098. This one-utility requirement is consistent with *Cross*, which held that “[w]hen a properly claimed invention meets at least one stated objective, utility under §101 is clearly shown” *Cross*, 753 F.2d at 1046 fn9, citing *Raytheon Co. v. Roper Corp.* 724 F.2d 951, 958 (Fed. Cir. 1983), *cert. denied*, 469 U.S. 835 (1984). Thus, the Examiner’s utility rejection depends on the invalidity of each of Applicants’ asserted uses. However, as the Examiner noted (at page 3 of the Office Action mailed February 12, 2001 (paper 11)), “applicants do indeed provide multiple well-established and specific utilities for a GPCR.” Inexplicably, the Examiner now states (at page 5 of the Examiner’s Answer mailed August 28, 2001 (paper 18)) that “since there was no specific and substantial asserted utility or a well-established utility for the claimed nucleic acids and encoded proteins, credibility of the utility was not assessed.”

The PTO guidelines state, “[a] rejection based on lack of utility should not be maintained if an asserted utility for the claimed invention would be considered specific, substantial, and credible by a person of ordinary skill in the art in view of all evidence of record.” 66 Fed. Reg. 1098. “Credibility is assessed from the perspective of one of ordinary skill in the art in view of the disclosure....” 66 Fed. Reg. 1098. As the Examiner noted (at page 3 of the Office Action mailed Feb. 12, 2001 (paper 11)), Applicants “do indeed provide multiple well-established and

specific utilities for a GPCR,” and one of ordinary skill in the art would agree with the Examiner that the present invention satisfies the utility standard.

The PTO utility examination guidelines also state,

[w]here the asserted utility is not specific or substantial, a *prima facie* showing [of no specific and substantial credible utility] must establish that it is more likely than not that a person of ordinary skill in the art would not consider that any utility asserted by the Applicants would be specific and substantial. The *prima facie* showing must contain the following elements: (1) An explanation that clearly sets forth the reasoning used in concluding that the asserted utility for the claimed is not both specific and substantial nor well-established; (2) Support for factual findings relied upon in reaching this conclusion; and (3) An evaluation of all relevant evidence of record, including utilities taught in the closest prior art.

(66 Fed. Reg. 1098). Further, “[o]ffice personnel are reminded that they must treat as true a statement of fact made by Applicants in relation to an asserted utility, unless countervailing evidence can be provided that shows that one of ordinary skill in the art would have a legitimate basis to doubt the credibility of such a statement” (66 Fed. Reg. 1098-99).

This provision is consistent with the case law. See, *In re Gazave*, 379 F.2d 973 (C.C.P.A. 1967) (finding that the utility standard was met where “appellant’s assertions of usefulness in his specification appear to be believable on their face and straightforward, at least in the absence of reason or authority in variance”); *Ex parte Dash*, 27 U.S.P.Q.2d 1481, 1484 (Bd. Pat. App. & Int’l 1993) (holding that “[a] disclosure of a utility satisfies the utility requirement of section 101 unless there are reasons for the artisan to question the truth of such disclosure.”) Similarly, in *In re Jolles*, claims to pharmaceutical compounds and methods of use were rejected under §101 and §112. The court held, “it is proper for the examiner to ask for substantiating evidence unless one with ordinary skill in the art would accept the allegations as obviously correct” (628 F.2d 1322, 1327 (C.C.P.A. 1980)). See also, *In re Brana*, 51 F.3d 1560, 1563 (Fed. Cir. 1995) (stating that “[o]nly after the PTO provides evidence showing that one of ordinary skill in the art would reasonably doubt the asserted utility does the burden shift to the Applicants to provide rebuttal evidence sufficient to convince such a person of the invention’s asserted utility,” and holding that the PTO did not meet this burden.)

In re: Glucksmann, et al.

Appl. No.: 09/464,685

Filing Date: 12/16/99

Page 21

In the present case, the utility rejection has not been supported in the required manner. As discussed above, the Examiner's objections are not properly grounded in the authority cited and are in fact inconsistent with practices in the art. Accordingly, the Examiner has not made a *prima facie* showing of no utility and the rejection should be withdrawn.

CONCLUSION

In view of the arguments presented above, Applicants contend that each of claims 73, 74, 81, and 88-96 is patentable. Therefore, reversal of the rejections under 35 U.S.C. § 101 and 35 U.S.C. § 112, first paragraph, is respectfully solicited.

It is not believed that extensions of time or fees for net addition of claims are required, beyond those, which may otherwise be provided for in documents accompanying this paper. However, in the event that additional extensions of time are necessary to allow consideration of this paper, such extensions are hereby petitioned under 37 CFR § 1.136(a), and any fee required therefore (including fees for net addition of claims) is hereby authorized to be charged to Deposit Account No. 16-0605.

Respectfully submitted,

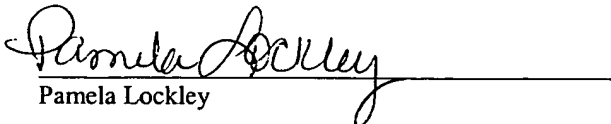


Leigh W. Thorne
Registration No. 47,992

CUSTOMER NO. 00826
ALSTON & BIRD LLP
Bank of America Plaza
101 South Tryon Street, Suite 4000
Charlotte, NC 28280-4000
Tel Raleigh Office (919) 862-2200
Fax Raleigh Office (919) 862-2260

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as first class mail in an envelope addressed to: Commissioner for Patents, Washington, DC 20231 on October 29, 2001.



Pamela Lockley

APPEALED CLAIMS

73. A method for detecting the presence of a polypeptide having an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence shown in SEQ ID NO:1; and
- (b) the amino acid sequence encoded by the cDNA contained in ATCC

Deposit No. PTA-2369;

said method comprising contacting the sample with a compound which selectively binds to any one of the polypeptides of (a) – (b) and determining whether the compound binds to said polypeptides in the sample.

74. The method of claim 73, wherein the compound which binds to the polypeptide is an antibody.

81. A method for modulating the activity of a polypeptide having an amino acid sequence selected from the group consisting of:

- (a) the amino acid sequence shown in SEQ ID NO:1; and
- (b) the amino acid sequence encoded by the cDNA contained in ATCC

Deposit No. PTA-2369;

said method comprising contacting any one of polypeptides (a) – (b) or a cell expressing any one of polypeptides (a) – (b) with a compound which binds to the polypeptide in a sufficient concentration to modulate the activity of the polypeptides.

88. A method for screening a cell to identify an agent that binds with a polypeptide having an amino acid sequence shown in SEQ ID NO:1 in said cell, said method comprising contacting said cell with an agent and detecting an interaction between said polypeptide and agent.

89. A method for screening a cell to identify an agent that modulates the expression level or activity of the polypeptide having an amino acid sequence shown in SEQ ID NO:1 in said cell, said method comprising contacting said cell with an agent and detecting an interaction between said polypeptide and agent.

90. The method of claim 89, wherein said cell is a blood cell.

91. The method of claim 90, wherein said blood cell is a myeloid progenitor cell.

92. The method of claim 91, wherein said myeloid progenitor cell is a CD34⁺ cell.

93. The method of claim 89, wherein said agent increases the level or activity of said polypeptide.

94. The method of claim 89, wherein said agent decreases the level or activity of said polypeptide.

95. A method for assessing G-protein receptor expression in disease states of a patient, comprising contacting a tissue of said patient with an isolated antibody that selectively binds to the polypeptide shown in SEQ ID NO:1.

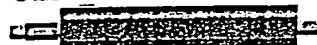
96. The method of claim 95, wherein the G-protein coupled receptor expression is involved in signal transduction.

[Pfam 6.2 \(St. Louis\)](#) : [Home](#) | [Analyze a sequence](#) | [Browse alignments](#) | [Text search](#) | [Swisspfam](#) | [Help](#)

Domain structure of proteins in the 7tm_1 Seed alignment

Pfam domains are large boxes. Small three-colored boxes are Pfam-B clusters. Mouseover to see domain descriptions. Click on box to enter family page. (Javascript is used for mouseover functionality.)

OLFJ_HUMAN P30954 OLFACTORY RECEPTOR-LIKE PROTEIN HGMP07J.



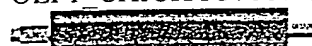
OL15_MOUSE P23275 OLFACTORY RECEPTOR 15 (OR3).



OLF6_RAT P23267 OLFACTORY RECEPTOR-LIKE PROTEIN F6.



OLF1_CHICK P37067 OLFACTORY RECEPTOR-LIKE PROTEIN COR1.



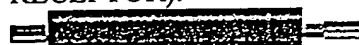
GU27_RAT P34987 GUSTATORY RECEPTOR GUST27.



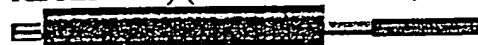
RTA_RAT P23749 PROBABLE G PROTEIN-COUPLED RECEPTOR RTA.



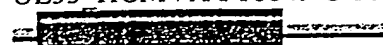
TA2R_HUMAN P21731 THROMBOXANE A2 RECEPTOR (TXA2-R) (PROSTANOID TP RECEPTOR).



PE24_HUMAN P35408 PROSTAGLANDIN E2 RECEPTOR, EP4 SUBTYPE (PROSTANOID EP4 RECEPTOR) (PGERECEPTOR, EP4 SUBTYPE).



UL33_HCMVA P16849 G-PROTEIN COUPLED RECEPTOR HOMOLOG UL33.



OPSB_HUMAN P03999 BLUE-SENSITIVE OPSIN (BLUE CONE PHOTORECEPTOR)

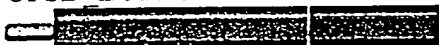
PIGMENT).



OPS3_DROME P04950 OPSIN RH3 (INNER R7 PHOTORECEPTOR CELLS OPSIN).



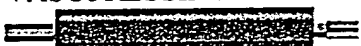
OPSD_LOLFO P24603 RHODOPSIN.



OPS1_DROME P06002 OPSIN RH1 (OUTER R1-R6 PHOTORECEPTOR CELLS OPSIN).



V2R_HUMAN P30518 VASOPRESSIN V2 RECEPTOR (RENAL-TYPE ARGININE VASOPRESSIN RECEPTOR)(ANTIDIURETIC HORMONE RECEPTOR) (AVPR V2).



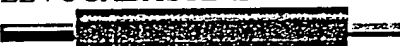
FSHR_BOVIN P35376 FOLLICLE STIMULATING HORMONE RECEPTOR PRECURSOR (FSH-R) (FOLLITROPINRECEPTOR).



TRFR_HUMAN P34981 THYROTROPIN-RELEASING HORMONE RECEPTOR (TRH-R) (THYROLIBERINRECEPTOR).



NTR1_HUMAN P30989 NEUROTENSIN RECEPTOR TYPE 1 (NT-R-1) (HIGH-AFFINITY LEVOCABASTINE-INSENSITIVE NEUROTENSIN RECEPTOR) (NTRH).



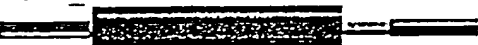
NY1R_HUMAN P25929 NEUROPEPTIDE Y RECEPTOR TYPE 1 (NPY1-R).



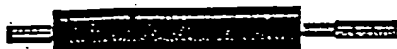
NYR_DROME P25931 NEUROPEPTIDE Y RECEPTOR (NPY-R) (PR4 RECEPTOR).



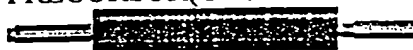
TLR1_DROME P30974 TACHYKININ-LIKE PEPTIDES RECEPTOR 86C (NKD).



NK1R_CAVPO P30547 SUBSTANCE-P RECEPTOR (SPR) (NK-1 RECEPTOR) (NK-1R).



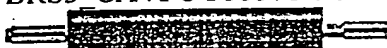
GPCR1_MOUSE P30731 PROBABLE G PROTEIN-COUPLED RECEPTOR FROM T-CELLS
PRECURSOR (GLUCOCORTICOID-INDUCED RECEPTOR).



CCKR_HUMAN P32238 CHOLECYSTOKININ TYPE A RECEPTOR (CCK-A RECEPTOR)
(CCK-AR).



BRS3_CAVPO P35371 BOMBESIN RECEPTOR SUBTYPE-3 (BRS-3).



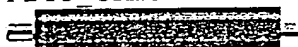
PAFR_CAVPO P21556 PLATELET ACTIVATING FACTOR RECEPTOR (PAF-R).



THRR_CRILO Q00991 THROMBIN RECEPTOR PRECURSOR.



P2Y5_CHICK P32250 P2Y PURINOCEPTOR 5 (P2Y5) (PURINERGIC RECEPTOR 5) (6H1).



EBI2_HUMAN P32249 EBV-INDUCED G PROTEIN-COUPLED RECEPTOR 2 (EBI2).



US28_HCMVA P09704 G-PROTEIN COUPLED RECEPTOR HOMOLOG US28 (HHRF3).



US27_HCMVA P09703 G-PROTEIN COUPLED RECEPTOR HOMOLOG US27 (HHRF2).



C5AR_CANFA P30992 C5A ANAPHYLATOXIN CHEMOTACTIC RECEPTOR (C5A-R).



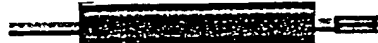
RDC1_CANFA P11613 G PROTEIN-COUPLED RECEPTOR RDC1.



G10D_RAT P31392 PROBABLE G PROTEIN-COUPLED RECEPTOR G10D (NOW).



SSR1_HUMAN P30872 SOMATOSTATIN RECEPTOR TYPE 1 (SS1R) (SRIF-2).



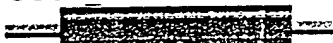
OPRD_MOUSE P32300 DELTA-TYPE OPIOID RECEPTOR (DOR-1) (K56) (MSL-2).



APJ_HUMAN P35414 PROBABLE G PROTEIN-COUPLED RECEPTOR APJ.



GUSB_BOVIN P35350 POSSIBLE GUSTATORY RECEPTOR TYPE B (PPR1 PROTEIN).



CKR7_HUMAN P32248 C-C CHEMOKINE RECEPTOR TYPE 7 PRECURSOR (C-C CKR-7) (CC-CKR-7) (CCR-7) (MIP-3 BETA RECEPTOR) (EBV-INDUCED G PROTEIN-COUPLED RECEPTOR).



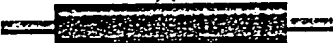
CX3I_RAT P35411 CX3C CHEMOKINE RECEPTOR 1 (C-X3-C CKR-1) (CX3CR1) (FRACTALKIN RECEPTOR) (GPR13) (RBS11).



CKR1_HUMAN P32246 C-C CHEMOKINE RECEPTOR TYPE 1 (C-C CKR-1) (CC-CKR-1) (CCR-1) (CCR1) (MACROPHAGE INFLAMMATORY PROTEIN-1 ALPHA RECEPTOR) (MIP-1A RECEPTOR).



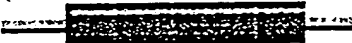
CCR4_BOVIN P25930 C-X-C CHEMOKINE RECEPTOR TYPE 4 (CXC-R4) (CXCR-4) (SDF-1 RECEPTOR) (STROMAL CELL-DERIVED FACTOR 1 RECEPTOR) (FUSIN) (LEUKOTAXIN RECEPTOR).



IL8A_HUMAN P25024 HIGH AFFINITY INTERLEUKIN-8 RECEPTOR A (IL-8R A) (IL-8 RECEPTOR TYPE 1) (CXCR-1) (CDW128).



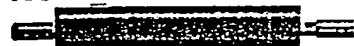
CCR5_HUMAN P32302 C-X-C CHEMOKINE RECEPTOR TYPE 5 (CXC-R5) (CXCR-5) (BURKITT'S LYMPHOMA RECEPTOR 1) (MONOCYTE-DERIVED RECEPTOR 15) (MDR15).



BRB2_HUMAN P30411 B2 BRADYKININ RECEPTOR (BK-2 RECEPTOR).



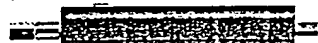
AG2R_BOVIN P25104 TYPE-1 ANGIOTENSIN II RECEPTOR (AT1).



AG22_MOUSE P35374 TYPE-2 ANGIOTENSIN II RECEPTOR (AT2).



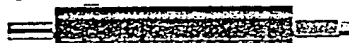
MC3R_MOUSE P33033 MELANOCORTIN-3 RECEPTOR (MC3-R).



EDG1_HUMAN P21453 PROBABLE G PROTEIN-COUPLED RECEPTOR EDG-1.



CB2R_HUMAN P34972 CANNABINOID RECEPTOR 2 (CB2) (CB-2) (CX5).



CB1R_HUMAN P21554 CANNABINOID RECEPTOR 1 (CB1) (CB-R) (CANN6).



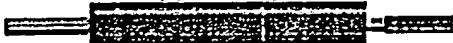
ACM1_HUMAN P11229 MUSCARINIC ACETYLCHOLINE RECEPTOR M1.



AA1R_BOVIN P28190 ADENOSINE A1 RECEPTOR.



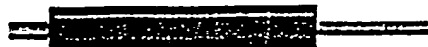
5H2A_CRIGR P18599 5-HYDROXYTRYPTAMINE 2A RECEPTOR (5-HT-2A) (SEROTONIN RECEPTOR)(5-HT-2).



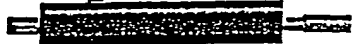
5H5A_MOUSE P30966 5-HYDROXYTRYPTAMINE 5A RECEPTOR (5-HT-5A) (SEROTONIN RECEPTOR)(5-HT-5).



5H6_RAT P31388 5-HYDROXYTRYPTAMINE 6 RECEPTOR (5-HT-6) (SEROTONIN RECEPTOR)(ST-B17).



HH2R_CANFA P17124 HISTAMINE H2 RECEPTOR (GASTRIC RECEPTOR I).



D2DR_BOVIN P20288 D(2) DOPAMINE RECEPTOR.



A1AD_HUMAN P25100 ALPHA-1D ADRENERGIC RECEPTOR (ALPHA 1D-ADRENOCEPTOR) (ALPHA-1AADRENERGIC RECEPTOR).



DADR_HUMAN P21728 D(1A) DOPAMINE RECEPTOR.



B1AR_HUMAN P08588 BETA-1 ADRENERGIC RECEPTOR.



5HT1_DROME P20905 5-HYDROXYTRYPTAMINE RECEPTOR 1 (5-HT RECEPTOR) (SEROTONIN RECEPTOR).



5H7_HUMAN P34969 5-HYDROXYTRYPTAMINE 7 RECEPTOR (5-HT-7) (5-HT-X) (SEROTONIN RECEPTOR)(5HT7).



5H1B_HUMAN P28222 5-HYDROXYTRYPTAMINE 1B RECEPTOR (5-HT-1B) (SEROTONIN RECEPTOR)(5-HT-1D-BETA) (S12).



5H1A_HUMAN P08908 5-HYDROXYTRYPTAMINE 1A RECEPTOR (5-HT-1A) (SEROTONIN RECEPTOR) (5-HT1A) (G-21).





Review

The quest to deduce protein function from sequence: the role of pattern databases

T.K. Attwood*

School of Biological Sciences, The University of Manchester, Oxford Road, Manchester M13 9PT, UK

Received 17 May 1999; accepted 3 August 1999

Abstract

In the wake of the numerous now-fruitful genome projects, we have witnessed a 'tsunami' of sequence data and with it the birth of the field of bioinformatics. Bioinformatics involves the application of information technology to the management and analysis of biological data. For many of us, this means that databases and their search tools have become an essential part of the research environment. However, the rate of sequence generation and the haphazard proliferation of databases have made it difficult to keep pace with developments, even for the cognoscenti. Moreover, increasing amounts of sequence information do not necessarily equate with an increase in knowledge, and in the panic to automate the route from raw data to biological insight, we may be generating and propagating innumerable errors in our precious databases. In the genome era upon us, researchers want rapid, easy-to-use, *reliable* tools for functional characterisation of newly determined sequences. For the pharmaceutical industry in particular, the Pandora's box of bioinformatics harbours an information-rich nugget, ripe with potential drug targets and possible new avenues for the development of therapeutic agents. This review outlines the current status of the major pattern databases now used routinely in the analysis of protein sequences. The review is divided into three main sections. In the first, commonly used terms are defined and the methods behind the databases are briefly described; in the second, the structure and content of the principal pattern databases are discussed; and in the final part, several alignment databases, which are frequently confused with pattern databases, are mentioned. For the new-comer, the array of resources, the range of methods behind them and the different tools required to search them can be confusing. The review therefore also briefly mentions a current international endeavour to integrate the diverse databases, which effort should facilitate sequence analysis in the future. This is particularly important for target-discovery programmes, where the challenge is to rationalise the enormous numbers of potential targets generated by sequence database searches. This problem may be addressed, at least in part, by reducing search outputs to the more focused and manageable subsets suggested by searches of integrated groups of family-specific pattern databases. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Bioinformatics; Similarity search; Sequence alignment; Pattern recognition; Function annotation

Abbreviations: HMM, hidden Markov model; ICGEB, International Center for Genetic Engineering and Biotechnology; ISREC, Institut Suisse de Recherche Experimentale sur le Cancer; MIPS, Munich Information Centre for Protein Sequences; PAM, point accepted mutation; PDB, Protein DataBank; PIR, protein identification resource; PSD, protein sequence database; SRS, sequence retrieval system.

* Tel.: +44-161-275-5766; fax: +44-161-275-5082.

1. Introduction

Ten years from the dawn of the field of bioinformatics, we are harvesting the abundant fruits of a variety of genome projects and, in spite of early flood warnings, the resultant torrent of sequence information has all but broken our databanks. Biological databases are now a central part of the research environment, but many have evolved simply as a by-product of a particular individual's research project, with no thought that they might one day become valuable international treasures. Consequently, some have not stood the test of time (most do not survive beyond the first five years [1]). Others are creaking under the strain of information overload, their underlying technologies never having been designed to cope with such volumes of data. Still others have managed to survive via collaborative efforts, some on an international scale. The protein sequence database (PSD), for example, evolved in the early 1960s from Margaret Dayhoff's research on the evolutionary relationships among proteins [2]. By 1980, the collection had grown to (a mere) 200 sequences [3], which in the last two decades has increased more than 600 fold to ~131,000 (release 61, June 1999). The PSD is now maintained collaboratively by PIR-International [4] and is one of the most comprehensive protein sequence collections currently available.

Today, there are hundreds of databanks around the world housing information at the levels of the genome, the proteome and even the *métabolome* [5]. The endeavour to cope with and rationalise these vast quantities of data has required global co-operation and ever increasing levels of automation in data handling and analysis. However, automation carries a price. In the field of genomics, for example, although software robots are essential to the process of functional annotation of newly determined sequences, they pose a threat to information quality because they can introduce and propagate mis-annotations [6]. The curators are aware of this and always strive to improve the quality of their resources, but databases are nevertheless historical products (or,

in some cases, historical accidents!) and are therefore far from perfect. To get the most from current biological databases it is thus important to have an understanding both of their powers and of their pitfalls.

The first step towards functional characterisation of a new sequence usually involves trawling a sequence database with tools such as BLAST [7] or FASTA [8]. Such searches quickly reveal similarities between the query and a range of database sequences. The trick then lies in the reliable inference of homology (the verification of a divergent evolutionary relationship) and, from this, the inference of function. Ideally, a search output will show unequivocal similarity to a well-characterised protein over the full length of the query. At worst, an output will reveal no significant hits, but the usual scenario is a list of partial matches to diverse proteins, many of them uncharacterised and some with dubious or contradictory annotations [9].

There are various reasons for this sort of confusion. For example, the increasing size of sequence databases and their population by greater numbers of poorer quality partial sequences (such as expressed sequence tags), gives rise to an increasing likelihood that high-scoring matches will be made to a query simply by chance. So-called low-complexity matches, in particular, may swamp search outputs — these are regions within a sequence that have high densities of particular residues (e.g. poly-GxP, such as occurs in repetitive, often tightly structured sequences like collagen; or polyglutamine tracts that occur in Huntingdon's disease protein; and so on). Although mechanisms are available for masking such sequences, their incautious use may also create complications. The modular and/or domain nature of many proteins also causes problems on different levels. First, when matching multidomain proteins, it may not be clear which domain or domains correctly correspond to the query. Second, even if the right domain has been identified, it may not be appropriate to transfer the functional annotation to the query because the function of the matched domain may be different, depending on its precise biological con-

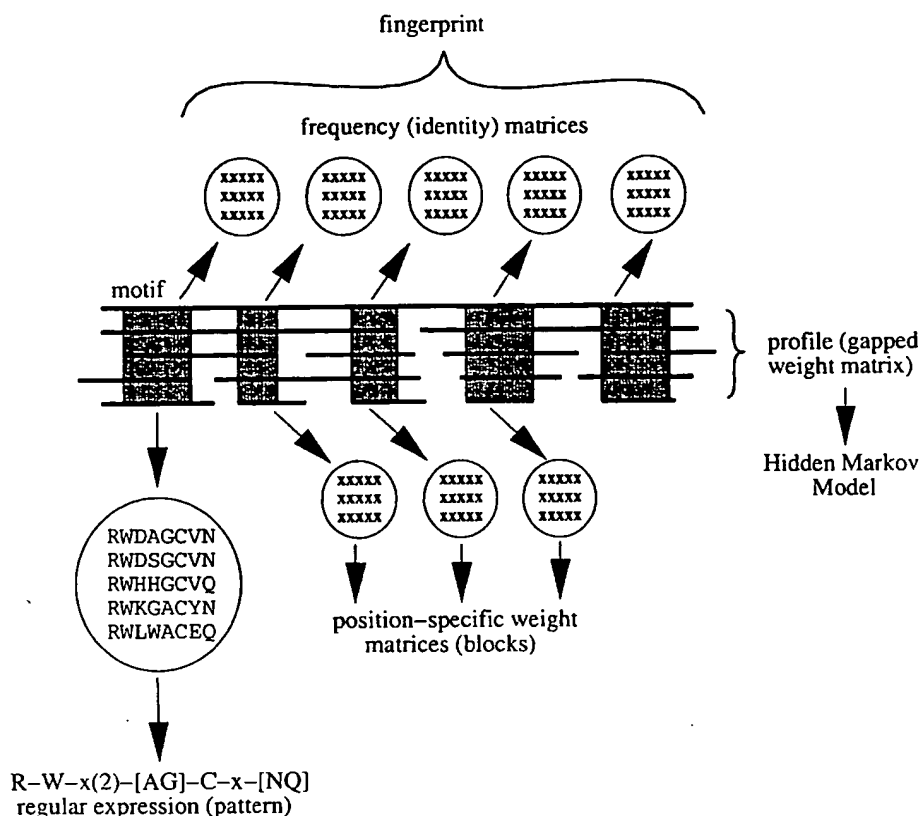


Fig. 1. At the heart of sequence analysis methods is the multiple sequence alignment. Application of these methods involves the derivation of some kind of representation of conserved features of the alignment, which may be diagnostic of structure or function. Various terms are used to describe the different types of data representation, as shown. Within a single conserved region (motif), the sequence information may be reduced to a single consensus expression (a regular expression), often simply referred to as a pattern. In this example, square brackets indicate residues that are allowed at this position of the motif and x denotes any residue, the (2) indicating that any residue can occupy *consecutive* positions in the motif. The term used to describe groups of motifs in which all the residue information is retained within a set of frequency (identity) matrices is a fingerprint. Adding a scoring scheme to such sets of frequency matrices results in position-specific weight matrices, or blocks. Using information from extended conserved regions that include gaps (usually referred to as domains) gives rise to profiles; and probabilistic models derived from alignment profiles are termed hidden Markov models.

text. Similar issues arise with the existence of multigene families, because database search techniques cannot differentiate between a matched orthologue (the functional counterpart of a sequence in another species) and a matched paralogue (a homologue that performs different but related functions within the same organism).

Achieving consistent, reliable functional assignments can be a complicated process. As a result, in addition to routine searches of the sequence databases, it is now customary to extend search

strategies to include a range of 'value-added' or pattern databases. These distil information within groups of related sequences into potent descriptors or discriminators that aid family diagnosis. Searching pattern databases is more sensitive and selective than sequence database searching because derived family discriminators can detect weaker regions of similarity. Different analytical approaches have been used to create a bewildering array of discriminators, which are variously termed regular expressions, rules, profiles, signa-

Table 1

Web addresses of pattern and alignment databases in common use; for a more exhaustive list, refer to the annual database issue of *Nucleic Acids Research* (<http://www3.oup.co.uk/nar/>)

PROSITE	http://www.expasy.ch/prosite/
BLOCKS	http://www.blocks.fhcrc.org/
PRINTS	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/IDENTIFY
http://	dna.Stanford.EDU/identify/
Profiles	http://www.isrec.isb-sib.ch/software/PFSCAN_form.html
Pfam	http://www.sanger.ac.uk/Software/Pfam/
ProDom	http://www.toulouse.inra.fr/prodom.html
SBASE	http://www.icgeb.trieste.it/sbase/
PIR-ALN	http://www-nbrf.georgetown.edu/pirwww/search/textpiraln.html
PROT-FAM	http://vms.mips.biochem.mpg.de/mips/programs/classification.html
DOMO	http://www.infobiogen.fr/~gracy/domo/
ProClass	http://pir.georgetown.edu/gfserver/proclass.html
ProtoMap	http://www.protomap.cs.huji.ac.il/
PIMA	http://dot.imgen.bcm.tmc.edu:9331/seq-search/protein-search.html
ProWeb	http://www.proweb.org/kinesin/ProWeb.html

tures, fingerprints, blocks, etc. [10] — these terms are summarised in Fig. 1. The different descriptors have different diagnostic strengths and weaknesses and different areas of optimum application and have been used to generate different pattern databases, which also tend to differ in content! The aim of this review is to provide an overview of the current status of pattern and alignment databases in common use and to provide pointers on how best to use them. As this is a rapidly developing area, a list of Web addresses is given in Table 1 to allow readers to obtain the most up-to-date information on the resources discussed.

2. The methods behind the databases

At the heart of the analysis methods that underpin pattern databases is the multiple sequence alignment. When building an alignment, as more distantly related sequences are included, insertions are often required to bring equivalent

parts of adjacent sequences into the correct register, as illustrated schematically in Fig. 1. As a result of this gap insertion process, islands of conservation emerge from a backdrop of mutational change. These conserved regions (typically around 10–20 amino acids in length) tend to correspond to the core structural or functional elements of the protein; they are most commonly termed motifs, but are also referred to as blocks, segments or features.

Several techniques have evolved to exploit the conservation encoded in sequence alignments, as shown in Fig. 2 [11]. Broadly, the methods fall into three categories, depending on whether they use single motifs, multiple motifs or full domain alignments. Whatever the approach, all involve the derivation of some kind of discriminatory representation of the conserved alignment elements — essentially, the conserved motifs provide a characteristic signature or fingerprint for the family, which can be used to facilitate diagnosis of future query sequences.

The diagnostic success of the different methods depends on how reliably true family members (true positives) can be distinguished from non-family members (true negatives). In practice, there is a crucial balance between the number of incorrect matches that are made (false positives) and the number of correct matches that are missed (false negatives) at a given scoring threshold. As shown in Fig. 3, for a given search, the distribution of true positive matches must be resolved from that of the true negatives, such that the overlap between them is minimised or eliminated. This is important because, for matches in the overlapping area, it can be difficult or impossible to determine which are correct (statistical approaches are used to assign confidence levels to matches in this area, but mathematical significance does not give biological proof). The different analytical methods that have been designed to improve the resolving power of database searches are outlined below.

2.1. Single-motif methods

Of the various approaches, single-motif (regular expression pattern) methods are easiest to

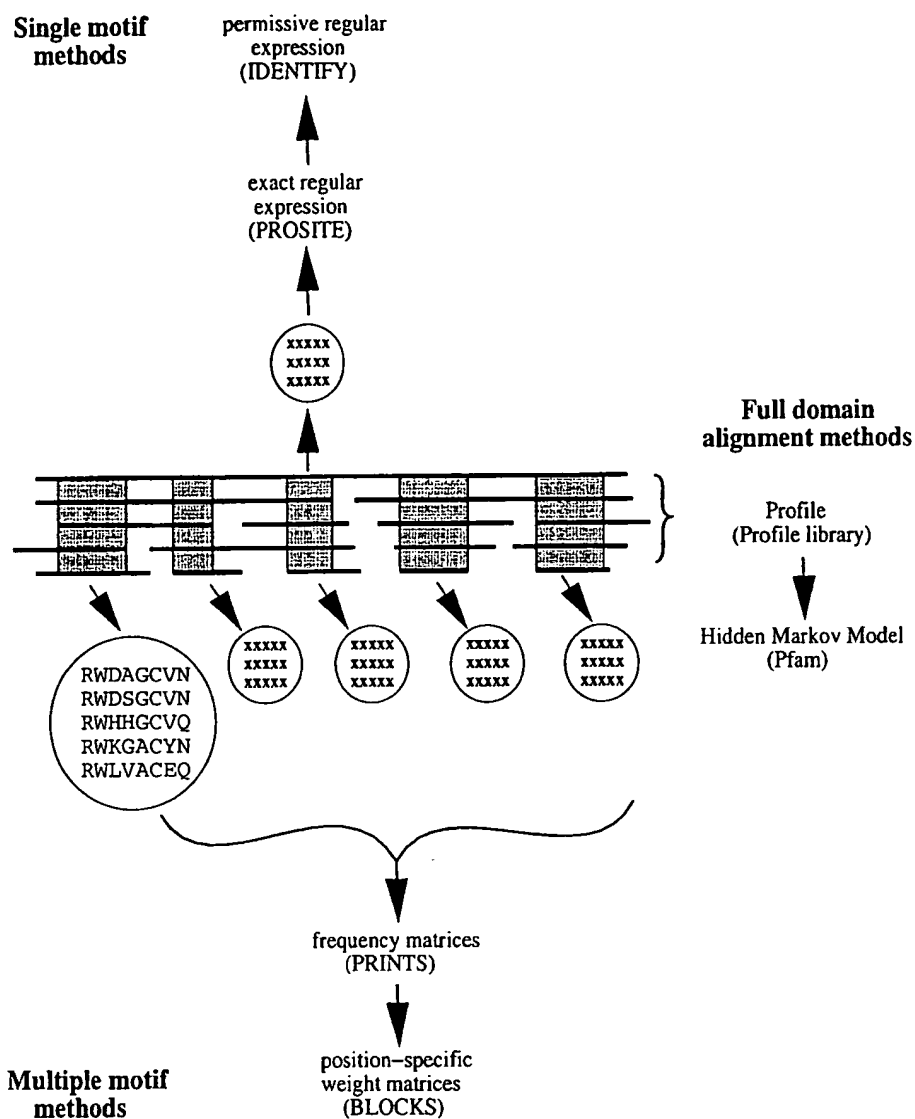


Fig. 2. Illustration of the three principal methods for building pattern databases: i.e. using single motifs, multiple motifs and full domain alignments. Single-motif (regular expression pattern) approaches have given rise to the PROSITE and IDENTIFY databases; multiple-motif methods have spawned the BLOCKS and PRINTS databases; and domain alignment methods have resulted in the Profiles and Pfam resources.

understand. The idea is that a particular protein family can be characterised by the single most conserved, often functionally important, region (e.g. an enzyme active site) observed in a sequence alignment. The motif is reduced to a consensus expression in which all but the most

significant residue information is discarded. For example, the short expression D-[ALV]-x-{YW}-T means that a conserved aspartic acid (D) residue is followed by a hydrophobic residue, which may be alanine (A), leucine (L) or valine (V); this is followed by an arbitrary residue (x) and any

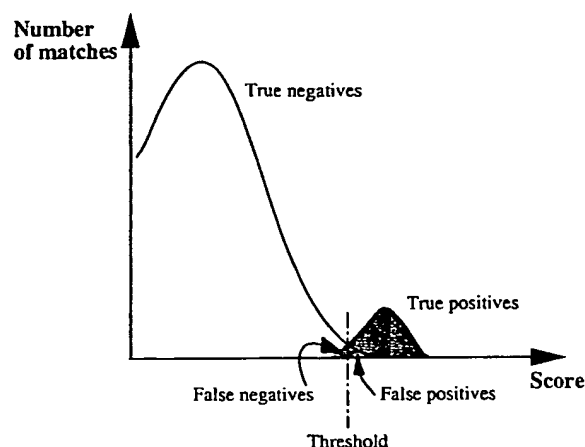


Fig. 3. Resolving true and false matches. In a database search, the desire is to establish which sequences are related to the query (i.e. are true positive) and which are unrelated (true negative). At a given scoring threshold, it is likely that several unrelated sequences will match erroneously (so-called false positives) and several correct matches will fail to be diagnosed (false negatives). In sequence analysis, the challenge is to improve diagnostic performance by capturing all (or the majority) of true positive family members, including no (or few) false positives and minimising or precluding false negatives.

residue *except* tyrosine (Y) or tryptophan (W); and finally a conserved threonine (T). No other residues or residue combinations are tolerated by the expression; matches to it must therefore be exact, or will be disregarded.

So rigid is this syntax that regular expression patterns do not perform well when used to represent highly divergent protein families. For example, these patterns will fail to match significant sequences if they contain a *single* amino acid difference — hence, the sequence DAMYT is a mis-match, in spite of matching the above expression in all but one position (it has a forbidden tyrosine as its fourth residue). Conversely, a pattern will match *anything* that corresponds to it exactly, regardless of whether it is a true family member. The problem is that matches to single motifs lack biological context — a match to a pattern is just a match to a pattern and may well only be fortuitous. To assess the likelihood of a match being 'real', it must be verified with corroborating evidence, whether via other database searches, the literature, experiment, etc.

Table 2

Overlapping sets of amino acids and their properties; these are used to create the permissive regular expressions used as the basis of the IDENTIFY resource

Residue property	Residue groups
Small	Ala, Gly
Small hydroxyl	Ser, Thr
Basic	Lys, Arg
Aromatic	Phe, Tyr, Trp
Basic	His, Lys, Arg
Small hydrophobic	Val, Leu, Ile
Medium hydrophobic	Val, Leu, Ile, Met
Acidic/amide	Asp, Glu, Asn, Gln
Small/polar	Ala, Gly, Ser, Thr, Pro

An approach that addresses the strict nature of exact regular expression matching is to assign amino acid residues to distinct, but overlapping, substitution groups corresponding to various biochemical properties (e.g. charge and size), as shown in Table 2. This is a biologically sensible approach because each amino acid has several properties and can serve different functions, depending on its biochemical context [12]. However, although the technique is more flexible, its inherent permissiveness brings with it an inevitable signal-to-noise trade-off — i.e. resulting patterns not only have the potential to make more true positive matches, but they will consequently also match more false positives. For example, the sequence DAMPS, which would be excluded by the exact regular expression above, would be matched by the permissive one (because Ser and Thr belong to the same group), even if threonine were biologically mandatory at the last position of the motif.

2.2. Multiple-motif methods

In response to the problems inherent in single-motif methods, diagnostic techniques subsequently evolved to exploit multiple motifs. Within a sequence alignment, it is usual to find not one, but several motifs that characterise the aligned family. Diagnostically, it makes sense to use many or all such conserved regions to build a family signature or fingerprint. In a database

search, there is then a greater chance of identifying a distant relative, whether or not all parts of the signature are matched. For example, a sequence that matches only four of seven motifs may still be diagnosed as a true match if the motifs are matched in the correct order in the sequence and the distances between them are consistent with those expected of true neighbouring motifs. The ability to tolerate mis-matches, both at the level of individual residues within motifs and at the level of motifs within the complete signature, renders multiple-motif matching a powerful diagnostic approach.

Different multiple-motif methods have arisen, depending both on the technique used to detect the motifs and on the scoring method employed. Probably the simplest to understand is the technique of fingerprinting [13]. Here, groups of conserved motifs are excised from a sequence alignment and used to create a series of frequency (identity) matrices — no mutation or other similarity data are used to weight the results. The scoring scheme is thus based on the calculation of residue frequencies for each position in the motifs, summing the scores of identical residues for each position of a retrieved match. However, the main strength of this approach also gives rise to its main weakness. In other words, because the method exploits observed residue frequencies, the scoring matrices are sparse and thus perform cleanly (with little noise) and with high specificity; at the same time, their absolute scoring potential is limited by the nature of the observed data. For richly populated families, this is not a problem, because the resulting matrices will reflect the constituent sequence diversity; but for poorly populated families, the matrices may be too sparse and may not encode sufficient variation to be able to detect distant relatives reliably, if at all.

One way to address this problem is to use mutation or substitution matrices to weight nonidentical residue matches. Commonly used scoring matrices include the PAM [14] and BLOSUM series [15]. The former is based on the concept of the point accepted mutation (PAM). PAM 250 is often used as a default matrix in comparison programs because it gives similarity scores equivalent

to 20% matches remaining between two sequences, the twilight zone [16] of similarity. The BLOSUM matrices, which are derived from observed substitutions in blocks of aligned sequences from the BLOCKS database, were designed to detect distant similarities more reliably than the Dayhoff series, which can only *infer* remote relationships because their substitution rates were derived from sets of highly similar sequences. Whatever the approach used, however, similarity matrices are inherently noisy because they indiscriminately weight both random matches and weak signals. Thus care should be taken to select a scoring matrix appropriate to the evolutionary distance at which relationships are being sought. For practical purposes, this means using a range of different matrices (though few people actually bother!).

2.3. Profile methods

An alternative philosophy to the motif-based approach of protein family characterisation adopts the principle that the variable regions between conserved motifs also contain valuable information. Here, the complete conserved portion of the alignment (including gaps) effectively becomes the discriminator. The discriminator, termed a profile, defines which residues are allowed at given positions, which positions are highly conserved and which degenerate, and which positions, or regions, can tolerate insertions. The scoring system is intricate and may include evolutionary weights and results from structural studies, as well as data implicit in the alignment. In addition, variable penalties may be specified to weight against insertions and deletions occurring within core secondary structure elements [17,18]. Profiles provide a sensitive means of detecting distant sequence relationships where only very few residues are well-conserved.

Just as there are different ways of using motifs to characterise protein families, so there are different ways of using domain alignments to build family discriminators. An extension of the concept of profiles lies in the application of hidden Markov models (HMMs) [19]. These are probabilistic models consisting of a number of

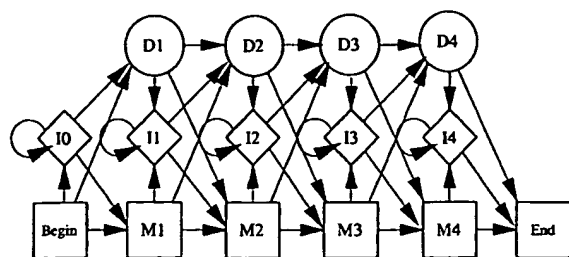


Fig. 4. Linear hidden Markov model (HMM). Each position of an alignment is represented as a match (M), an insert (I), or a delete (D) state in the HMM. This allows a query sequence to be aligned by assigning the most probable state transition to each of its residues.

interconnecting states — they are essentially linear chains of match, delete or insert states that attempt to encode the sequence conservation within aligned families. A match state is assigned to each conserved column in a sequence alignment; an insert state allows for insertions relative to the match states; and delete states allow positions to be skipped. Probabilities or costs (negative log probabilities) are associated with each omission and each transition between states. To align a sequence is to find the highest-probability (lowest-cost) path through the HMM. A linear HMM is depicted in Fig. 4.

Although capable of providing precise descriptors for particular families, as with all methods, there are drawbacks. One problem arises from the specificity of profiles and HMMs. For example, they may be well trained for a given family, but an outlier that was not included in the training set may be missed if features of its sequence are incompatible with the model. Another problem relates to the automatic, iterative nature of HMM training; without adequate supervision, the process may include false positive matches, which may ultimately corrupt the model and lead to profile dilution.

3. Pattern databases

As a consequence of the range of sources of sequence data and the variety of ways of analysing sequences and encoding protein families, a

Table 3

Some of the major pattern databases in common use; in each case, the primary source is noted, together with the type of pattern stored (e.g. regular expression, fingerprint, HMM, etc.)

Pattern database	Data source	Stored information
PROSITE	SWISS-PROT	regular expressions (patterns)
PRINTS	SWISS-PROT/ TrEMBL	aligned motifs (fingerprints)
Profiles	SWISS-PROT	gapped weight matrices (profiles)
Pfam	SWISS-PROT/ TrEMBL	gapped domain alignments (HMMs)
BLOCKS	PROSITE/ PRINTS	aligned motifs (blocks)
IDENTIFY	BLOCKS/ PRINTS	permissive regular expressions (patterns)

number of different pattern databases have evolved to house the different descriptors outlined in the previous section. The databases and their associated methods are summarised in Table 3. Despite their differences, pattern databases have arisen from a common principle: i.e. homologous sequences share conserved motifs, presumably crucial to the structure or function of the protein, which can be used to build discriminators for particular protein families. An unknown query sequence may be searched against a library of such descriptors to determine whether or not it contains any of the predefined characteristics and hence whether or not it can be assigned to a known family. If the structure and function of the family is known, searches of pattern databases thus theoretically offer a fast track to the inference of biological function. Because these resources are derived from multiple sequence information, searches of them are often better able to identify distant relationships than are searches of the sequence databases. However, none of the pattern databases is yet complete. They should therefore be used to augment sequence searches, rather than to replace them. The status of some of the commonly used pattern resources is outlined below.

PROSITE, the first pattern database to have been developed, houses motifs in the form of

```

ID  BACTERIAL_OPSIN_1; PATTERN.
AC  PS00950;
DT  JUN-1994 (CREATED); JUN-1994 (DATA UPDATE); JUL-1998 (INFO UPDATE).
DE  Bacterial rhodopsins signature 1.
PA  R-Y-x-[DT]-W-x-[LIVMF]-[ST]-T-P-[LIVM](3).
NR  /RELEASE=36,74019;
NR  /TOTAL=22(22); /POSITIVE=22(22); /UNKNOWN=0(0); /FALSE_POS=0(0);
NR  /FALSE_NEG=1; /PARTIAL=1;
CC  /TAXO-RANGE=A????; /MAX-REPEAT=1;
DR  P19585, BAC1_HALS1, T; P29563, BAC2_HALS2, T; P96787, BAC3_HALSD, T;
DR  Q48334, BAC3_HALVA, T; P33970, BACH_HALHM, T; Q48315, BACH_HALHP, T;
DR  Q48314, BACH_HALHS, T; P16102, BACH_HALSP, T; P33742, BACH_HALSS, T;
DR  P94853, BACH_HALVA, T; P15647, BACH_NATPH, T; Q57101, BACR_HALAR, T;
DR  P02945, BACR_HALHA, T; P33969, BACR_HALHM, T; P33971, BACR_HALHP, T;
DR  P33972, BACR_HALHS, T; Q53496, BACR_HALSR, T; P94854, BACR_HALVA, T;
DR  P25964, BACS_HALHA, T; P33743, BACS_HALSS, T; P71411, BACT_HALSA, T;
DR  P42196, BACT_NATPH, T;
DR  Q53461, BACH_HALAR, P;
DR  P42197, BACT_HALVA, N;
3D  1BRD; 2BRD; 1BAC; 1BAD; 1BHA; 1BHB; 1BCT; 1SR1;
DO  PDOC00291;
//

```

Fig. 5. Example PROSITE entry, showing the data file for the bacteriorhodopsin pattern. When viewing PROSITE on the Web, accession numbers are hyperlinked, allowing direct access to the corresponding SWISS-PROT entry for each sequence matched. Similarly, the documentation file for a given pattern can be accessed via the hyperlinked PDOC accession number at the bottom of the file.

regular expressions [20]. The process of deriving regular expressions first requires the construction of a multiple alignment and then location of the conserved regions. The most conserved segment is selected and its sequence information reduced to a consensus pattern, which is used to search SWISS-PROT [21]. Results are checked manually to determine how well the pattern has performed — ideally, there should be only true matches. Patterns whose diagnostic performance is compromised by matching too many false positives are manually adjusted and SWISS-PROT is rescanned. The process of fine tuning is repeated until an optimal pattern is created. If a family cannot be fully characterised by a single motif, additional patterns are designed to encode other well-conserved parts of the alignment. The fine-tuning process is then repeated until a set of patterns is achieved that is capable of capturing all, or most, of the family without matching too many, or any, false positives. When the best pattern, or *set* of patterns, has been achieved, the results are manually annotated for inclusion in the database.

Entries are deposited in PROSITE in two distinct files: (i) a structured data file that houses the pattern and lists all matches in the parent

version of SWISS-PROT, as shown in Fig. 5; (ii) a free-format documentation or annotation file, which provides details of the characterised family and, where known, a description of the biological role of the chosen motif/s and a supporting bibliography. A number of features of the data file are worthy of note. Apart from the identifier (ID) and description (DE) lines, which identify the characterised family, aspects of the DR and especially the NR lines are crucial to understand. The DR lines list all true (T), possible (P), false (F) and missed/negative (N) matches to the pattern, which results are summarised in the NR lines. In the example shown in Fig. 3, 22 matches are made to the pattern, all of which are true, one is possible (a fragment) and there is a single false negative match, BACT_HALVA. Inspection of its sequence (e.g. by following its hyperlinked accession number, P42197, from this page on the Web) reveals that a disallowed asparagine in the penultimate position of the motif (RYVDWLLTTPLN^V) is the reason for the mismatch. Referring back to the pattern line, we see that only members of the group [LIVM] are allowed in the last three positions of the motif (as denoted by [LIVM](3)). The quality of a pattern can thus immediately be ascertained from

the NR lines, which are therefore probably the most important lines to inspect when first viewing a PROSITE entry. In some cases, there are numerous false positives and false negatives (especially for large super-families with substantial numbers of divergent sequences, such as G-protein-coupled receptors, lipocalins, etc.). Such patterns are diagnostically unreliable and are a limitation to the diagnostic potential of the database. PROSITE release 15 (July 1998), with updates to April 1999, contains 1014 entries characterised by 1352 patterns. The database is accessible for searching via the ExPASy Web server and is maintained collaboratively at the Swiss Institute of Bioinformatics.

BLOCKS, one of the first multiple-motif databases, is based on families already identified in PROSITE [22]. Here, motifs are detected automatically, using first, a modification of an algorithm that initially locates three conserved amino acids [23] and second, a motif-finding algorithm that searches for the highest scoring set of blocks that occur in the correct order without overlapping. Blocks found by both methods are considered reliable and are calibrated against SWISS-PROT to obtain a measure of the likelihood of a chance match. The calibrated blocks are then concatenated into the BLOCKS database. An indication of the diagnostic power of a block is given in terms of a strength value — strong blocks are more effective than weak blocks (strength less than 1100) at separating true positives from true negatives. In searching the database, however, more important than the strength of individual blocks is the *number* of blocks matched. High-scoring matches to individual blocks seldom have biological significance; conversely, matches to sets of blocks from the same family are unlikely to have arisen by chance (provided they match in the correct order with appropriate distances between them) and a probability value is calculated to reflect that likelihood. Release 11.0 contains 4034 blocks, representing 994 groups from PROSITE 15.

Recently, several other BLOCKS databases have been made available. For example, in BLOCKS+, supplementing the entries derived from PROSITE are blocks from families in

PRINTS that are not already in BLOCKS and then successively, any additional blocks from Pfam, ProDom and DOMO. BLOCKS+ is thus comprehensive, containing 9498 blocks from 2129 sequence groups. Complementing this resource is a version of PRINTS in which block-scoring methods have been exploited [22]. PRINTS' motifs tend to be deeper than those in BLOCKS because its source database is larger; the diagnostic performance of entries in the two resources can therefore differ, BLOCKS-format-PRINTS tending to be more prone to problems of noise. Because the BLOCKS databases are derived automatically, their entries are not annotated, but links are made to the corresponding PROSITE and PRINTS documentation files. The databases are accessible for searching via the Web server at the Fred Hutchinson Cancer Research Center in Seattle.

PRINTS, another of the early responses to the diagnostic limitations of regular expression matching, is based on the method of fingerprinting [24]. This approach uses groups of conserved motifs to build diagnostic signatures of family membership. The process involves manual creation of a seed alignment, and location and excision of conserved motifs for searching SWISS-PROT and TrEMBL. Results are examined to determine which sequences have matched all the motifs in the fingerprint; if there are more matches than were in the initial alignment, the additional information from these new sequences is added to the motifs and the database is searched again. This iterative process is repeated until no further complete fingerprint matches can be identified. The results are then annotated manually (with descriptions of the family, details of the structural or functional relevance of the motifs where known, cross-references to related databases, bibliographic references, etc.) prior to inclusion in the database.

Fingerprint diagnostic performance is indicated via a summary that lists how many sequences matched all the motifs and how many made only partial matches (i.e. failed to match one or more motifs). The fewer the partial matches, the better the fingerprint. The full potency of the method derives from the mutual context provided by

motif neighbours. The more motifs in a fingerprint, the better able it is to identify distant relatives, even when parts of the signature are absent; conversely, the fewer the motifs, the poorer the diagnostic performance. Fingerprints with only two motifs are diagnostically little better than single motifs and are therefore more likely to make false positive matches. When searching PRINTS, probability and expect values are calculated to assign a measure of confidence to both complete and partial matches.

Within PRINTS, motifs are encoded as ungapped, un-weighted local alignments. An important consequence of storing the motifs in this 'raw' form is that, unlike with regular expressions or other abstractions, no sequence information is lost. Different scoring methods may thus be superposed onto the motifs, conferring different scoring potentials, and hence different perspectives, on the same data. PRINTS may therefore provide the raw material for other pattern databases. PRINTS release 23.0 (June 1999) contains 1160 entries (6938 motifs), currently making it the most comprehensive manually annotated pattern database. The database is accessible for searching via the Web server in the School of Biological Sciences at the University of Manchester.

IDENTIFY is derived automatically from motifs in BLOCKS and PRINTS [12]. The program used to create the database constructs consensus expressions from the motifs, adopting a permissive approach in which different residues are tolerated according to a set of prescribed groupings (Table 2). These groups correspond to various biochemical properties, theoretically ensuring that the resulting expressions have sensible biochemical interpretations. However, as mentioned earlier, in practice this approach may lead to an increase in noise. When searching the resource, different levels of stringency are therefore offered from which to infer the significance of matches, rendering the approach diagnostically more powerful than exact pattern matching (which only offers match/no-match diagnoses). *IDENTIFY* is accessible from the Web server in the Department of Biochemistry at the University of Stanford.

Profiles are discriminators distilled from sequence information in complete domain alignments. As a result of their potency, they are used to complement some of the poorer regular expressions in PROSITE, or to provide a diagnostic alternative where extreme sequence divergence renders the use of regular expressions inappropriate. A compendium of profiles has been created at the Swiss Institute for Experimental Cancer Research (ISREC) in Lausanne. Each profile has separate PROSITE-compatible data and documentation files. This allows results that have been validated and annotated to an appropriate standard to be made available as an integral part of PROSITE [20]. As before, diagnostic performance can be ascertained from the DR and NR lines. Profiles are less prone to make false matches than are regular expressions, but the numbers released via PROSITE are only small (48 in July 1998). Nevertheless, profiles that have not yet achieved the necessary standard of validation and annotation (241 to date) are available for searching via ISREC's Web server.

Pfam is a collection of HMMs for a range of protein domains [25]. The resource is based on two distinct classes of alignment: hand-edited seed alignments, which are deemed to be accurate; and an automatically clustered set derived from ProDom families. The seed alignments are used to build HMMs, to which sequences are automatically aligned to generate final full alignments. If the initial alignments do not produce diagnostically sound HMMs, the seed is improved and the gathering process iterated until a good result is achieved. The methods that ultimately generate the best full alignment may vary for different families, so the parameters are saved to allow results to be reproduced. The collection of seed and full alignments, coupled with minimal annotations (often no more than a description line), related database and literature cross-references and the HMMs themselves, constitute Pfam-A. All sequence domains that are not included in Pfam-A are automatically clustered and deposited in Pfam-B. Although the methods and parameters used to create the full automatic alignment are noted, no indication is given of the diagnostic performance of a given HMM. Direct

Table 4

Some of the major alignment databases; in each case, the primary source is noted, together with the level of information stored (i.e. whether domain, family or super-family alignments)

Alignment database	Primary source	Stored information
ProDom	SWISS-PROT	domains
SBASE	SWISS-PROT	domains
ProtoMap	SWISS-PROT	families
PIR-ALN	PIR	super-families, families and domains
PROT-FAM	PIR	super-families, families and domains
ProClass	SWISS-PROT/PIR	super-families, families and domains
DOMO	SWISS-PROT/PIR	domains and repeats
PIMA	Entrez	domains

visualisation of the final alignment is therefore probably the best indicator of how sound its HMM is likely to be. Pfam is accessible for searching via the Sanger Centre Web server; release 4.1 (July 1999) encodes 1488 domains.

4. Alignment and family-related databases

In addition to the range of pattern resources described above, several alignment databases are also available for searching via the Web. The construction of alignment and pattern databases is based on different principles, so the two types of resource should not be confused. The main difference between them is that alignment databases tend to be derived simply by automatic clustering of sequence databases. This allows them to be more comprehensive than pattern resources, because they do not depend on manual crafting of family discriminators. However, searches of alignment databases are often less sensitive because they are usually based on implementations of BLAST. Some well-known alignment resources are listed in Table 4.

ProDom is an automatic compilation of 'hom-

ologous' domains [26] created via a procedure based on PSI-BLAST [7]. Version 99.1 contains 44,345 domains with at least 2 sequences, of which 2652 are linked to the Protein DataBank (PDB) [27]. A recent addition to the resource is ProDom-CG, a compendium of domains built from complete genome data. The database is accessible for interrogation with the Sequence Retrieval System (SRS) [28] and for BLAST searching via the Web server of the Institut National de la Recherche Agronomique. Emphasis has been placed on the graphical user interface, which facilitates analysis of protein relationships. However, being automatically derived, no annotations or validations are provided and although links are made to the PDB for ~5% of entries, these are generic links from the constituent sequences rather than from the domains themselves. Discovering the biological meaning of domains can thus be difficult, involving extensive cross-checking with other resources.

SBASE is a library of domain sequences derived from structural and functional segments annotated in SWISS-PROT, PIR or the literature [29]. Entries are grouped on the basis of standard names and further classified on the basis of BLAST similarity. The resource, which was developed to assist domain recognition, is maintained collaboratively by the International Center for Genetic Engineering and Biotechnology (ICGEB), Trieste, Italy and the ABC Institute for Biochemistry and Protein Research, Gödöllő, Hungary. *SBASE* is accessible for BLAST searching via the ICGEB Web server; version 6.0 (October 1998) contains 1038 groups.

ProtoMap classifies sequences in SWISS-PROT into groups of related proteins [30]. Clustering is effected at different levels of confidence, resulting in a hierarchical organisation that divides the sequences into well-defined groups, which mostly correlate with biological families and superfamilies. The resource was designed to help reveal relationships between families and to facilitate the detection of sub-families. ProtoMap release 2.0 (July 1998) provides a classification of 72,623 sequences. The resource is accessible for searching via the Hebrew University Web server.

PIR-ALN is a database of annotated protein sequence alignments derived automatically from the PIR-International PSD at the National Biomedical Research Foundation in Washington [31]. The database includes alignments at super-family, family and so-called homology domain levels. Sequences are grouped in the same super-family if they are similar from end to end; super-families are further subdivided into families containing sequences that are 45% identical; and segments corresponding to the same domain in two or more super-families are the basis of domain alignments. All domain alignments are deposited in the DOMAINDB database, which is used to screen new sequences for already defined domains. The March 1999 release of PIR-ALN contains 3983 alignments, including 1480 super-family and 371 domain alignments. The resource can be queried with the ATLAS information retrieval system at the PIR Web site.

PROT-FAM is based on an automatic clustering of the PIR-International PSD at the Munich Information Center for Protein Sequences (MIPS) [32]. Sequences that share 50% identity are clustered into families, and families are further clustered into super-families if they share ~30% identity. Sequences are assigned to the same family if they are similar from N- to C-terminus, while regions showing ~30% identity that do not cover the full sequence length are annotated as domains. Domains are deposited into the HOMDOM database, which is used to characterise new sequences by means of the pre-defined domains. For all families, super-families and domains that contain more than one sequence, alignments are created using PILEUP [33]. The September 1998 release of PROT-FAM included 6000 families with two sequences and ~6500 families containing three or more; ~3800 super-families derived from more than one family; and 361 domains. These are available for querying via the MIPS Web site.

ProClass is a value-added database built upon the PIR-International PSD, PROSITE and SWISS-PROT [34]. It organises nonredundant SWISS-PROT and PIR sequences according to relationships defined collectively by PIR super-families and PROSITE patterns. By combining

global similarities and motifs into a single classification scheme, ProClass was designed to facilitate identification of domain and family relationships, and classification of multidomain proteins. ProClass release 4.0 (September 1998) contains 122,253 sequence entries, ~60% of which are classified into ~3500 families. The resource is available for searching from the PIR Web server.

DOMO is a database of 'homologous' domain alignments computed automatically from a non-redundant amalgam of SWISS-PROT and PIR [35]. The domains have been compiled in FASTA format to permit fast searching using BLAST and sequence alignment using CLUSTALW [36]. The resource was designed as an aid to determine domain arrangements, their evolutionary relationships and their key conserved amino acids. DOMO can be queried via SRS at the Infobiogen Web site. Release 1.2 (April 1998) contains 99,058 domains clustered into 8877 sequence alignments. Query results are linked to other databases to provide complementary information on related proteins and their families. Where 3D structures of representative sequences are known, links to the atomic coordinates and structure classification resources are provided. If the domain structure is unknown, pointers are given to a composite secondary structure prediction obtained from a variety of different techniques. As with other automatically generated resources, the structure links are generic and do not relate directly to the domains themselves; understanding their biological significance can therefore be difficult.

PIMA is a collection of conserved motifs generated by clustering the NCBI's Entrez database [37]. For families of two or more sequences, alignments are created using the pattern-induced multiple alignment program [38] and these are scanned for the presence of conserved regions. If an alignment contains one or more such elements, additional alignments are created by excision of these conserved segments. Currently, the PIMA database includes 22,416 alignments, each of which contributes a single pattern to the resource; it is available for searching with modified versions of FASTA via the Baylor College of

Medicine Search Launcher Web pages. Here, another database has been created by extracting the locations of all annotated domains and sites from sequences contained in the Entrez, PROSITE, BLOCKS and PRINTS databases. The BEAUTY utility incorporates this information directly into BLAST search results [39]; for each match, a schematic display allows direct comparison of the locations of conserved regions.

5. Which database is best?

The plethora of available databases presents bewildering choices to the would-be sequence analyst. Which is diagnostically most reliable? Which has the most useful annotations? Which is the most comprehensive? Which should I use? At first sight, the alignment resources appear to be the most comprehensive. But they are largely based on automatic clustering of sequence databases and their search tools thus tend to involve flavours of BLAST or FASTA, which are less sensitive than searches of family-specific patterns. It is difficult to assess the quality of particular resources and it would be invidious to try. Each has different diagnostic strengths and weaknesses, each offers different family coverage and different levels of annotation — each has certain merits and demerits. Nevertheless, some general points bear consideration.

Automatically generated databases carry no annotations. The advantage of searching them is that they are more comprehensive than their manually derived counterparts. The disadvantage is that there may be no way to ascertain the biological significance of a match, if indeed it has any (that a match has been made does not mean an evolutionary relationship necessarily exists). This is important to understand in light of resources that house ‘homology domains’ — automatic methods detect *similarities*, but it is for the user to *infer* homology from supporting biological evidence. Related issues arise in resources that calculate evolutionary trees from their automatically created alignments; if levels of stringency are sufficiently high, alignments and their trees may be sound; but at low stringency, results

are likely to be error prone and relationships should be inferred with caution.

Amongst pattern databases, single-motif methods that rely on exact regular expression pattern-matching have diagnostic limitations; such methods tolerate no similarity, so will fail to diagnose sequences that contain subtle changes not catered for by the pattern. Moreover, single motifs offer no biological context within which to assess the significance of a match. Multiple-motif approaches inherently offer improved diagnostic reliability by virtue of the mutual context provided by motif neighbours. Thus, if a query fails to match all the motifs in a signature, the pattern of matches formed by the remaining motifs still allows the user to make a confident diagnosis.

Pattern resources derived from existing databases have the limitation that they offer no further family coverage. Nevertheless, they have the advantage of implementing different analytical methods from their source databases, thus offering different scoring potentials on the same data and furnishing important opportunities to diagnose relationships missed by the original implementations.

Finally, manually annotated databases are set apart from their automatically created counterparts by virtue of (i) providing *validation* of results and (ii) offering detailed information that helps to place conserved sequence information in structural or functional contexts. This is vital for the user, who not only wants to discover whether a sequence has matched a predefined motif, but also needs to understand its biological significance.

6. Composite pattern databases

If, today, comprehensive sequence analysis requires accessing a variety of disparate databases, gathering the range of different outputs and arriving at some sort of consensus view of the results, in the future this process should become more straightforward. The curators of PROSITE, Profiles, PRINTS, Pfam and ProDom are currently creating a unified database of protein families, termed InterPro. The aim is to pro-

vide a single family annotation resource, based on existing documentations in PROSITE and PRINTS and on the minimal annotations provided in Pfam. Each InterPro family will link to different entries in its satellite pattern databases. This will simplify sequence analysis for the user, who will thereby have access to a central resource for protein family diagnosis.

This effort is supported by the curators of the BLOCKS databases, who, realising the problems associated with providing detailed family documentation, are developing a dedicated protein family Web site, termed proWeb [40]. This facility provides information about individual families via links to existing Web resources maintained by researchers in their own fields. ProWeb can facilitate the task of annotators by providing convenient access to family information and obviating the need for annotators themselves to become 'expert' on all proteins.

7. Conclusion

Creating and searching pattern databases are activities that lie at different ends of a fallible chain of events. We begin with a sequence alignment, we create some kind of scoring function to encode the conservation within the alignment (a scoring matrix, HMM, etc.), we store the discriminators in a database and we search them with different algorithms. Problems arise if unrelated sequences have crept into the alignment, which in turn lead to errors in the discriminators, which then give ambiguous or incorrect search results. Alternatively, the discriminators may be sound, but the search algorithms may not be sufficiently sensitive to allow unequivocal diagnosis, leading the user to false conclusions of family ties. If the user has performed this experiment on a newly determined sequence and submits the results to one of the sequence databases, the annotation error becomes available for mass propagation.

Recently, there has been doom-mongering in the literature about the quality of our databases, some harbingers of misfortune predicting a future error catastrophe. At the same time, claims of success for some approaches to family classifi-

cation and function prediction have been equally overdone. A more balanced view recognises that our databases and search routines are not perfect, but with the right approach we can avoid the pitfalls of jumping to over-pessimistic or over-zealous conclusions.

Until we have sufficient experimental data available, pattern and sequence databases are probably the best tools we have for accessing the functional and evolutionary clues latent in the sequences flooding from the genome projects. Pattern databases offer several benefits: (i) by distilling multiple sequence information into family descriptors, trivial errors in the underlying sequences may be diluted; (ii) annotation errors may be quickly spotted if the description of one sequence differs from that of its family; and (iii) they allow specific diagnoses, placing individual sequences in a family context for a more informed assessment of possible function. By contrast, searches of sequence databases tend to reveal only *generic* similarities, making precise pinpointing of a particular biological niche more difficult.

While there is some overlap between them, the contents of the pattern databases differ. Together they encode ~2000 families, including globular and membrane proteins, modular polypeptides and so on. It has been estimated that the total number of families might be in the range 1000 to 10,000, so there is a long way to go before any of the databases can be considered complete. Thus, in building a search strategy, it is good practice to include all available pattern resources, to ensure that the analysis is as comprehensive as possible and that it takes advantage of a variety of search methods. Where there is consensus, diagnoses can be made with greater confidence.

Unfortunately, creating and annotating family descriptors is time-consuming, so pattern databases have not kept pace with the deluge of sequence data. Consequently, by comparison with the sequence repositories, they are still very small. Nevertheless, as they become more comprehensive, as the volume of sequence data expands and search outputs become more complex, their diagnostic potency ensures that pattern databases will play an increasingly

important role as the post-genome quest to assign functional information to raw sequence data gains pace.

Acknowledgements

I am grateful to the Royal Society for a University Research Fellowship.

References

- [1] L.B.M. Ellis, D. Kalumbi, Financing a future for public biological data, *Bioinformatics*, in press.
- [2] M.O. Dayhoff, R.V. Eck, M.A. Chang, M.R. Sochard, *Atlas of Protein Sequence and Structure*, Vol. 1, National Biomedical Research Foundation, Silver Spring, MD, 1965.
- [3] M.O. Dayhoff, R.M. Schwartz, H.R. Chen, L.T. Hunt, W.C. Barker, B.C. Orcutt, *Nucleic acid sequence bank*, *Science* 209 (1980) 1182.
- [4] W.C. Barker, J.S. Garavelli, P.B. McGarvey, C.R. Marzec, B.C. Orcutt, G.Y. Srinivasarao, L.-S.L. Yeh, R.S. Ledley, H.W. Mewes, F. Keiffer, A. Tsugita, C. Wu, *The PIR-International protein sequence database*, *Nucleic Acids Research* 27 (1) (1999) 39–43.
- [5] S.G. Oliver, M.K. Winson, D.B. Kell, F. Baganza, *Systematic functional analysis of the yeast genome*, *Trends in Biotechnology* 16 (1998) 373–378.
- [6] T. Doerks, A. Bairoch, P. Bork, *Protein annotation: detective work for function prediction*, *Trends in Genetics* 14 (1998) 248–250.
- [7] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*, *Nucleic Acids Research* 25 (17) (1997) 3389–3402.
- [8] W.R. Pearson, *Empirical statistical estimates for sequence similarity searches*, *Journal of Molecular Biology* 276 (1) (1998) 71–84.
- [9] K. Hoffmann, *Protein classification and functional assignment*, in: *Trends Guide to Bioinformatics*, Elsevier, 1998, pp. 18–21.
- [10] T.K. Attwood, *Exploring the language of bioinformatics*, in: H. Stanbury (Ed.), *Oxford Dictionary of Biochemistry and Molecular Biology*, Oxford University Press, 1997, pp. 715–723.
- [11] T.K. Attwood, D.J. Parry-Smith, *Introduction to Bioinformatics*, Addison-Wesley/Longman, Harlow, Essex, UK, 1999.
- [12] C.G. Nevill-Manning, T.D. Wu, D.L. Brutlag, *Highly specific protein sequence motifs for genome analysis*, *Proceedings of the National Academy Sciences of the USA* 95 (1998) 5865–5871.
- [13] D.J. Parry-Smith, T.K. Attwood, *ADSP: a new package for computational sequence analysis*, *CABIOS* 8 (5) (1992) 451–459.
- [14] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, *A model of evolutionary change in proteins*, in: M.O. Dayhoff (Ed.), *Atlas of Protein Sequence and Structure*, vol. 5 (suppl. 3), National Biomedical Research Foundation, Washington, DC, 1978, pp. 345–352.
- [15] J.G. Henikoff, S. Henikoff, *Amino acid substitution matrices from protein blocks*, *Proceedings of the National Academy Sciences of the USA* 89 (1992) 10915–10919.
- [16] R.F. Doolittle, *Of URFs and ORFs: a Primer on How to Analyse Derived Amino Acid Sequences*, University Science Books, Mill Valley, CA, 1986.
- [17] M. Gribskov, A.D. McLachlan, D. Eisenberg, *Profile analysis: detection of distantly related proteins*, *Proceedings of the National Academy Sciences of the USA* 84 (13) (1987) 4355–4358.
- [18] R. Luthy, U. Xenarios, P. Bucher, *Improving the sensitivity of the sequence profile method*, *Protein Science* 3 (1) (1994) 139–146.
- [19] R. Hughey, A. Krogh, *Hidden Markov models for sequence analysis: extension and analysis of the basic method*, *CABIOS* 12 (2) (1996) 95–107.
- [20] K. Hoffmann, P. Bucher, L. Falquet, A. Bairoch, *The PROSITE database, its status in 1999*, *Nucleic Acids Research* 27 (1) (1999) 215–219.
- [21] A. Bairoch, R. Apweiler, *The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1999*, *Nucleic Acids Research* 27 (1) (1999) 49–54.
- [22] J.G. Henikoff, S. Henikoff, S. Pietrovskiy, *New features of the BLOCKS database servers*, *Nucleic Acids Research* 27 (1) (1999) 226–228.
- [23] H.O. Smith, T.M. Annau, S. Chandrasegaran, *Finding sequence motifs in groups of functionally related proteins*, *Proceedings of the National Academy Sciences of the USA* 87 (1990) 826–830.
- [24] T.K. Attwood, D.R. Flower, A.P. Lewis, J.E. Mabey, S.R. Morgan, P. Scordis, J. Selley, W. Wright, *PRINTS prepares for the new millennium*, *Nucleic Acids Research* 27 (1) (1999) 220–225.
- [25] A. Bateman, E. Birney, R. Durbin, S.R. Eddy, R.D. Finn, E.L.L. Sonhammer, *Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins*, *Nucleic Acids Research* 27 (1) (1999) 260–262.
- [26] J. Gouzy, F. Corpet, D. Kahn, *Recent improvements of the ProDom database of protein domain families*, *Nucleic Acids Research* 27 (1) (1999) 263–267.
- [27] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer, M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimonouchi, M. Tasurni, *The protein data bank: a computer based archival file for macromolecular structures*, *Journal of Molecular Biology* 112 (1977) 535–542.
- [28] T. Etzold, A. Ulyanov, P. Argos, *SRS: information-retrieval system for molecular-biology data-banks*, *Methods in Enzymology* 266 (1996) 114–128.

- [29] J. Murvai, K. Vlahovick, E. Barta, C. Szepesvári, C. Acatrinei, S. Pongor, The SBASE protein domain library, release 6.0: a collection of annotated protein sequence segments, *Nucleic Acids Research* 27 (1) (1999) 257–259.
- [30] G. Yona, N. Linial, N. Tishby, M. Linial, A map of the protein space: an automatic hierarchical classification of all protein sequences, in: *Proceedings of ISMB*, The AAI Press, Menlo Park, CA, 1998, pp. 212–221.
- [31] G.Y. Srinivasarao, L.-S.L. Yeh, C.R. Marzec, B.C. Orcutt, W.C. Barker, F. Pfeiffer, Database of protein sequence alignments: PIR-ALN, *Nucleic Acids Research* 27 (1) (1999) 284–285.
- [32] H.W. Mewes, K. Heumann, A. Kaps, K. Mayer, F. Pfeiffer, S. Stocker, D. Frishman, MIPS: a database for genomes and protein sequences, *Nucleic Acids Research* 27 (1) (1999) 44–48.
- [33] J. Devereux, P. Haeblerli, O. Smithies, A comprehensive set of sequence analysis programs for the VAX, *Nucleic Acids Research* 12 (1) (1984) 387–395.
- [34] C.H. Wu, S. Shivakumar, H. Huang, ProClass protein family database, *Nucleic Acids Research* 27 (1) (1999) 272–274.
- [35] J. Gracy, P. Argos, DOMO: a new database of aligned protein domains, *Trends in Biochemical Science* 23 (12) (1998) 495–497.
- [36] J.D. Thompson, D.G. Higgins, T. Gibson, CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Research* 22 (22) (1994) 4673–4680.
- [37] G.D. Schuler, J.A. Epstein, H. Ohkawa, L.A. Kans, Entrez: molecular biology database and retrieval system, *Methods in Enzymology* 266 (1996) 141–162.
- [38] R.F. Smith, T.F. Smith, Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling, *Protein Engineering* 5 (1) (1992) 35–41.
- [39] K.C. Worley, B.A. Wiese, R.F. Smith, BEAUTY: an enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results, *Genome Research* 5 (2) (1995) 173–184.
- [40] S. Henikoff, S.A. Endow, E.A. Greene, Connecting protein family resources using the proWeb network, *Trends in Biochemical Science* 21 (1996) 444–445.

atrophy), and it is likely that others have yet to be found. A search of the working draft sequence yielded 286 potential paralogs of the 971 known human disease genes with entries in OMIM and either SwissProt or TREMBL protein databases. A similar screen of 603 classic drug target proteins identified 18 new potential paralogs. Together, these groups offer an intriguing collection of candidates – genes that might cause related disorders when mutated or that might encode new targets for drug screens.

Our understanding of disease mechanisms might also lead to the identification of new therapeutic targets. Profiling gene-expression changes in biological systems that model disease might lead to the identification of pathways that play a crucial role in pathogenesis. Such an endeavor is currently under way for the polyglutamine expansion diseases. Furthermore, consistent genetic changes that occur in easily accessible tissues in model organisms might provide surrogate markers for drug screens. In addition, an understanding of the common polymorphisms that occur in drug target proteins might help predict which patients will respond appropriately to therapy.

What lies ahead?

With a bounty of information being served up, it is important to keep in mind both the many strengths and the limitations of the

current data set. The working draft of the human genome that is accessible in publicly available databases includes almost one billion base pairs of finished sequence. However, nearly 75% of BACs are unfinished, currently consisting of as many as 10–20 unassembled sequence fragments each. Unfinished, unassembled sequence presents difficulties during gene mapping, might contain contamination and might be inadvertently assembled to create artificial duplications or deletions. Over the coming year, it is hoped that full coverage (8–10-fold redundancy) will be achieved for clones spanning the entire physical map, followed shortly thereafter by finished sequence. At that point, >96% of the euchromatic human genome will be in the database. Closing the remaining gaps, which might contain biologically important information, will require screening additional large-insert libraries, a process that is anticipated to take until 2003. Finally, new techniques might be needed to close recalcitrant gaps and to generate sequence from heterochromatic regions that probably contain highly polymorphic tandem repeats.

While we eagerly await completion of the finished genome sequence, our ability to mine the information we seek is rapidly evolving. Gene prediction, or annotation, is much more difficult in humans than in the fly, worm or yeast as a result of the large size of the genome. New computer algorithms and high-throughput techniques for gene identification and verification will be needed. Comparisons

with the genomes of other vertebrates will probably speed up this process and might reveal conserved regulatory regions that control the expression of orthologous genes. The Mammalian Gene Collection Project aims to assemble a comprehensive collection of full-length human cDNAs, providing a valuable resource for those studying gene function. Furthermore, by extending the SNP data set to include all common variants, the identification of disease genes and genetic modifiers should be greatly facilitated. This hope – of using the genome to help define causes and cures for human disease – underlies much of the excitement surrounding the release of the working draft.

Acknowledgements

A.P.L. is a Howard Hughes Medical Institute physician postdoctoral fellow.

References

- 1 Venter, C.V. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
- 2 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

Andrew P. Lieberman

Imke Puls

Kenneth H. Fischbeck*

Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Building 10, Room 3B14, 10 Center Drive, MSC 1250, Bethesda, MD 20892, USA.

*e-mail: kf@ninds.nih.gov

Techniques & Applications

A compendium of specific motifs for diagnosing GPCR subtypes

Teresa K. Attwood

Analysis of G-protein-coupled receptor (GPCR) subtypes has attracted considerable interest because some drugs that act on GPCRs cause therapeutic problems as a result of their failure to differentiate between subtypes. In this article, an extensive compendium of diagnostic 'fingerprints' for GPCR subtypes and their families will be described. These fingerprints offer new opportunities to investigate correlations between specific sequence motifs and ligand binding or

G-protein coupling, and are likely to prove valuable both in seeking novel receptors in genome data and in the characterization of orphan receptors.

G-protein-coupled receptors (GPCRs) constitute a vast group of cell-surface proteins that includes hormone, neurotransmitter, growth factor, light and odorant receptors. Approximately 2000 members populate ~50 families within the rhodopsin-like superfamily, accounting for

~1% of the vertebrate genome¹. With so many GPCRs known, and perhaps hundreds awaiting discovery in the human genome, these receptors are of interest to the pharmaceutical industry because of the opportunities they afford for yielding novel drug targets^{1–4}.

More than 50% of prescription drugs act on GPCRs; however, some have efficacy problems and limiting side-effects because the compounds do not differentiate between receptor subtypes. There is thus

Box 1. Identification of GPCRs using pattern databases

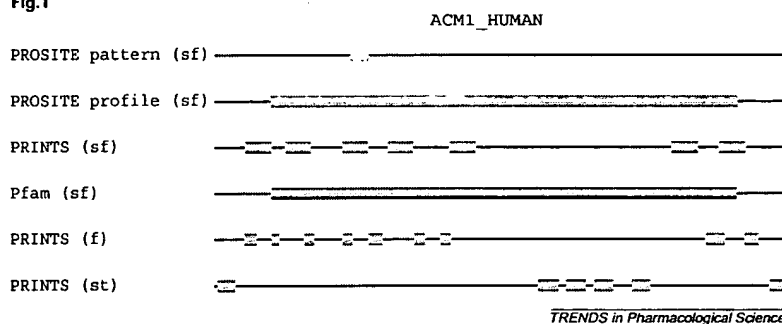
Protein pattern databases are becoming increasingly valuable as diagnostic resources that complement the ubiquitous sequence similarity search tool BLAST. Pattern databases house characteristic family signatures, which are encoded in different ways within the different resources: some encode single motifs (e.g. PROSITE patterns); others use groups of motifs in the form of fingerprints (e.g. PRINTS); and others encode virtually the full family alignment (e.g. PROSITE profiles and Pfam). Because the underlying analysis methods are different, inevitably the databases have different diagnostic strengths and weaknesses. It is therefore instructive to compare the results of searching a range of these resources using the same query sequence. A convenient way of doing this is to use the InterPro interface at <http://www.ebi.ac.uk/interpro/scan.html>. The graphical output in Fig. 1 shows the result of searching PROSITE patterns, PROSITE profiles, Pfam and PRINTS with the human muscarinic acetylcholine M₁ receptor, ACM1_HUMAN.

As shown, PROSITE patterns encode only single short motifs (yellow), whereas PROSITE profiles (orange) and Pfam (blue) utilize almost the complete sequence. By contrast, PRINTS fingerprints (green) encode groups of motifs that differentiate between regions of sequence that characterize the superfamily (sf) and those

that characterize the family (f) and receptor subtype (st). Thus, it is evident from the comparison that although PROSITE patterns, PROSITE profiles and Pfam only furnish superfamily diagnoses, PRINTS provides a more fine-grained result. The detail conferred by a fingerprint match lends PRINTS a significant part of its diagnostic power. Using PRINTS, we can see immediately that the superfamily fingerprint encodes seven motifs [hyperlinks to the database confirm that these are the transmembrane (TM) domains], whereas the family and receptor subtype fingerprints comprise different parts of the terminal, loop and TM regions. The mutual context of motif neighbours within a fingerprint offers a unique diagnostic advantage. By contrast with the 'pin-point' matches of PROSITE patterns and the 'blanket' matches of PROSITE

profiles and Pfam, PRINTS motifs explicitly capture, and map, functionally and structurally important biological features. This is valuable for several reasons: (1) in analyses of uncharacterized genome data, fingerprints are not limited to superfamily-level diagnoses, but provide sufficient depth to be able to pinpoint particular receptor subtypes, thereby facilitating the identification of novel receptors (M.D.R. Croning and T.K. Attwood, unpublished); (2) by storing motifs that differentiate between families and between receptor subtypes, correlations with specific residues involved in ligand-binding and G-protein coupling can be investigated; and hence (3) such fine-tuning, and the explicit encoding of motifs involved in ligand-binding, yields greater promise for our future ability to characterize orphan receptors.

Fig. 1



TRENDS in Pharmacological Sciences

considerable interest in attaining therapeutic selectivity by identifying the single receptor subtype that affects a particular physiology. The goal is to be able to design drugs without, or at least with less, side-effects, while retaining the desired function. Muscarinic agonists, for example, gained attention in research into Alzheimer's disease following the realization that the cardiovascular and gastrointestinal side-effects of nonselective muscarinic agonists could be avoided (i.e. muscarinic acetylcholine M₁ receptors in the brain might be involved in cognition, whereas other muscarinic receptor subtypes regulate heart and gastrointestinal functions⁵).

Identification of GPCRs

Routinely, computational strategies for identifying GPCRs tend to involve

searches of sequence databases [e.g. using standard tools such as BLAST (Ref. 6)] and sometimes also of so-called 'pattern' databases, which house diagnostic protein family 'signatures' (Box 1). However, it is apparent that BLAST 'sees' similarity between pairs of sequences in a rather limited way: it reveals generic similarities (e.g. it can show that the sequences being compared share several hydrophobic regions) but it cannot recognize individual family traits⁷ (i.e. it cannot distinguish the differences between the sequences, such as specific ligand-binding motifs). Similarly, most pattern databases tend to provide generic signatures that are only capable of diagnosing superfamily relationships. Thus, these databases might recognize that a sequence belongs to the rhodopsin-like GPCR superfamily, but they cannot offer insights into the

particular family to which it belongs. For researchers interested in, for example, the treatment of obesity and wishing specifically to identify type 4 melanocortin receptors (which are important in regulating appetite and body weight), a superfamily-level diagnosis is of limited value. Therefore, it seemed that it might be advantageous to develop a more fine-grained analytical approach for detecting GPCRs.

Identification of specific receptor subtypes

To facilitate the identification of particular subtypes, a systematic analysis of GPCRs was undertaken. Sequence alignments were created manually⁸ for each of the different superfamilies and for their families and receptor subtypes. Regions of similarity and differences between alignments

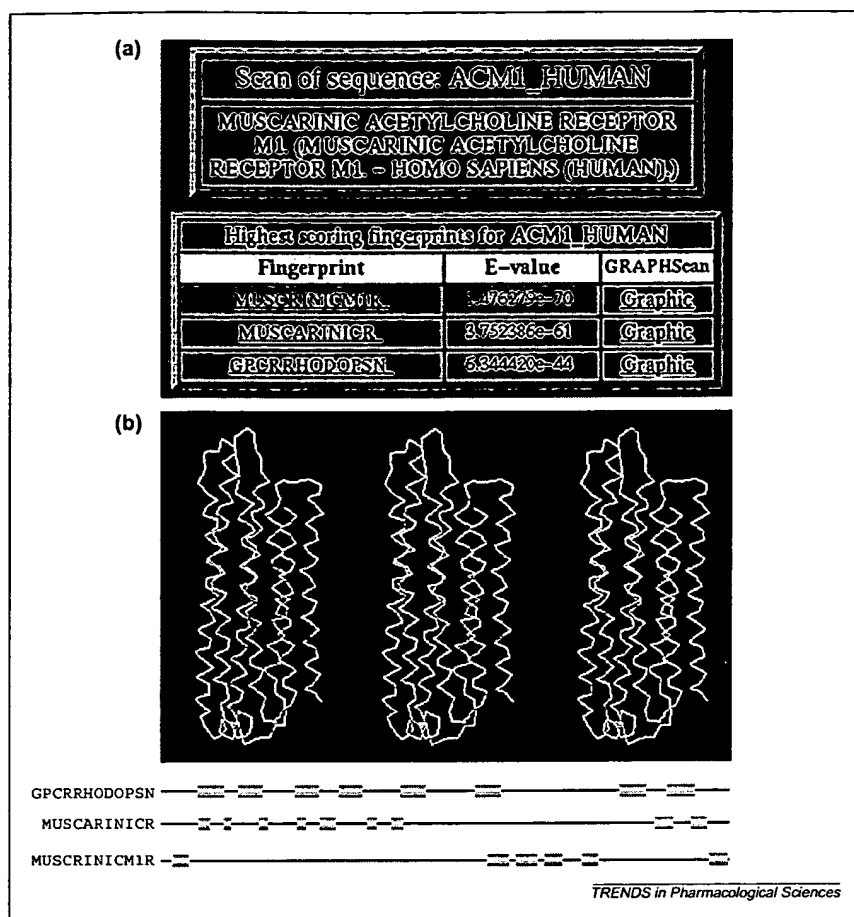


Fig. 1. (a) Hierarchical diagnosis returned from a PRINTS fingerprint search with the human muscarinic acetylcholine M₁ receptor, ACM1_HUMAN (the search is effected simply by pasting the full sequence, its identifier or its accession number into the Web form at <http://bioinf.man.ac.uk/cgi-bin/dbbrowser/fingerPRINTScan/muppetFPScan.cgi>). The result shows that three fingerprints have been matched, indicating that the sequence is likely to be a member of the rhodopsin-like G-protein-coupled receptor (GPCR) superfamily (fingerprint GPCRRHODOPSN), belonging to the muscarinic receptor family (MUSCARINICR) and being specifically an M₁ receptor subtype (MUSCRINICM1R). The E-values in the centre of the table provide the measure of confidence in the results (E-values indicate the number of matches one would expect to see by chance: the smaller the number, the more likely the matches are to be biologically meaningful). Here, the results are all statistically significant (i.e. above the threshold value of 10⁻⁴). (b) From left to right, the results in (a) are mapped in three dimensions onto a crude model and are illustrated schematically below. The coloured bars denote the relative locations and lengths of the constituent motifs within each fingerprint. The different regions that characterize the receptors at each level are clearly evident: motifs in the superfamily fingerprint encode each of the seven TM domains; those in the family fingerprint encode parts of TM and loop regions (here, TM domains 1, 3, 4, 5 and 7, the second cytoplasmic, and second and third external loops), the motifs mostly clustering around the ligand-binding domain; and motifs in the subtype fingerprint are drawn from the third cytoplasmic loop and the N- and C-terminal domains (not shown in 3D), areas known to be involved in regulating the selectivity and intensity of G-protein coupling¹.

were then located and used to build a range of discriminatory 'fingerprints'. Fingerprints are groups of conserved motifs that together provide a signature of family membership (motifs tend to reflect functionally or structurally important regions within a protein family [e.g. transmembrane (TM) domains, protein-protein interaction sites, ligand-binding sites, and so on], thereby characterizing the families in which they are found). For the purposes

of this analysis, within superfamilies the motifs encoded the only features common to all members (i.e. the scaffold of seven TM domains)^{9,10}. Conversely, at the family level, the motifs focused on those regions that characterized the particular family, but distinguished it from the parent superfamily; predictably, these were usually small parts of TM and loop regions. For receptor subtypes, the distinguishing traits were largely present in the N- and

C-terminal regions, and in the third cytoplasmic loop.

To date, >200 GPCR-specific fingerprints have been created and made available as an integral component of the PRINTS fingerprint database¹¹ (<http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/printscontents.html#Receptors>). By searching PRINTS with a given query, it is thus possible to make a hierarchical diagnosis, indicating to which superfamily and family the sequence belongs and which subtype it most resembles, as illustrated for the human M₁ receptor in Fig. 1a.

Biological significance of receptor motifs

To gain a deeper insight into the biological relevance of these database matches, the results can be rationalized in three dimensions by mapping the constituent motifs of the different fingerprints onto a crude model¹². For these purposes, an old model based on the structure of bacteriorhodopsin¹³ was used. Knowing that this was unlikely accurately to represent a GPCR (Ref. 14), our aim was simply to help visualize the relative three-dimensional (3D) locations of the motifs, rather than to ascertain precise atomic positions. As shown in Fig. 1b, the superfamily fingerprint encodes the 7TM scaffold, providing the architectural blueprint for all members; the family fingerprint focuses on the loop regions and on specific portions of the TM domains; and the subtype fingerprint is drawn from the third cytoplasmic loop and the N- and C-terminal domains. This is consistent with our expectation that portions of the TM segments are likely to constitute the ligand-binding domain, whereas the large intracellular region, unique to each subtype, is likely to constitute the effector-coupling domain¹⁵.

Similar results can be visualized for all the GPCR families housed in PRINTS, either using the fingerPRINTScan suite¹⁶ (Fig. 1a) or the BLAST PRINTS server¹⁷, both of which are accessible from the PRINTS home page (<http://www.bioinf.man.ac.uk/dbbrowser/PRINTS>). Alternatively, a powerful new resource that allows comparison of results from searches of PRINTS, PROSITE (Ref. 18) and Pfam¹⁹ is the integrated database of protein families, domains and functional sites known as InterPro (Ref. 20). By means of the graphical output from InterPro's sequence search, it is possible

to place the fingerprint matches in context and see at a glance which regions of a sequence are matched by the different resources. The example discussed in Box 1 demonstrates the fine-tuning that fingerprints add to the diagnostic process.

Concluding remarks

GPCR fingerprints allow specific diagnoses, from the level of the superfamily down to the individual receptor subtype. No other computational approach currently offers such a hierarchical discriminatory system for this important class of receptors. The resource is thus a valuable complement to family and domain databases such as PROSITE and Pfam, offering potent diagnostic opportunities that have not been realised by other pattern-recognition methods. Furthermore, fingerprint selectivity offers new opportunities to explore in more detail correlations between specific motifs and ligand binding or G-protein coupling. With the availability of the first draft of the human genome, this collection of diagnostic GPCR fingerprints promises to find application in computational strategies to identify potential new drug targets and to characterize orphan receptors.

Acknowledgements

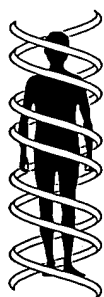
I am grateful to the Royal Society for a University Research Fellowship.

References

- 1 Bockaert, J. and Pin, J.P. (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J.* 18, 1723–1729
- 2 Marchese, A. *et al.* (1999) Novel GPCRs and their endogenous ligands: expanding the boundaries of physiology and pharmacology. *Trends Pharmacol. Sci.* 20, 370–375
- 3 Stadel, J.M. *et al.* (1997) Orphan G protein-coupled receptors: a neglected opportunity for pioneer drug discovery. *Trends Pharmacol. Sci.* 18, 430–437
- 4 Herz, J.M. *et al.* (1997) Molecular approaches to receptors as targets for drug discovery. *J. Recept. Signal Transduct. Res.* 17, 671–776
- 5 Sedlak, J. (1998) G-protein coupled receptors: research into receptor subtypes refines applications. *Genet. Eng. News* 18, 13
- 6 Altschul, S. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410
- 7 Wright, W. *et al.* (1999) BLAST PRINTS – an alternative perspective on sequence similarity. *Bioinformatics* 15, 523–524
- 8 Parry-Smith, D.J. *et al.* (1998) CINEMA – a novel colour interactive editor for multiple alignments. *Gene* 221, GC57–GC63
- 9 Attwood, T.K. and Findlay, J.B.C. (1993) Design of a discriminating fingerprint for G-protein-coupled receptors. *Protein Eng.* 6, 167–176
- 10 Attwood, T.K. and Findlay, J.B.C. (1994) Fingerprinting G-protein-coupled receptors. *Protein Eng.* 7, 195–203
- 11 Attwood, T.K. *et al.* (2000) PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res.* 28, 225–227
- 12 Findlay, J.B.C. *et al.* (1990) The structure of G-protein-linked receptors. *Biochem. Soc. Symp.* 56, 1–8
- 13 Henderson, R. and Schertler, G.F. (1990) The structure of bacteriorhodopsin and its relevance to the visual opsins and other seven-helix G-protein coupled receptors. *Philos. Trans. R. Soc. London Ser. B* 326, 379–389
- 14 Hibert, M.F. *et al.* (1993) This is not a G protein-coupled receptor. *Trends Pharmacol. Sci.* 14, 7–12
- 15 Peralta, E.G. *et al.* (1987) Distinct primary structures, ligand-binding properties and tissue-specific expression of four human muscarinic acetylcholine receptors. *EMBO J.* 6, 923–929
- 16 Scordis, P. *et al.* (1999) FingerPRINTScan: intelligent searching of the PRINTS motif database. *Bioinformatics* 15, 799–806
- 17 Wright, W. *et al.* (1999) BLAST PRINTS – an alternative perspective on sequence similarity. *Bioinformatics* 15, 523–524
- 18 Apweiler, R. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* 29, 37–40
- 19 Hofmann, K. *et al.* (1999) The PROSITE database, its status in 1999. *Nucleic Acids Res.* 27, 215–219
- 20 Bateman, A. *et al.* (2000) The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266

Teresa K. Attwood

School of Biological Sciences, The University of Manchester, Oxford Road, Manchester, UK M13 9PT.
e-mail: attwood@bioinf.man.ac.uk



What does the human genome sequence mean to you?

For a thorough and independent analysis of the importance of the February publications, don't miss the following articles in related titles:

Lawrence, R.N. (2001) Initial analyses of the human genome: does it reveal anything new? *Drug Discov. Today* 6(5), 221–222

Roses, A.D. (2001) The human medicine project has started: place your bets now. *Drug Discov. Today* 6(6), 282–284

Weinberg, R.A. (2001) A question of strategy. *Trends Biochem. Sci.* 26, 207–208

Lee, C. (2001) The incredible shrinking Human Genome. *Trends Genet.* 17, 187–188

Charlesworth, D. *et al.* (2001) Genome sequences and evolutionary biology, a two-way interaction. *Trends Ecol. Evol.* 16, 235–242

Commentaries relating to the publication of the draft sequences can be found both in print, in our sister *Trends* titles, and online in the Commentary section of BioMedNet (updated daily at <http://news.bmn.com/commentary>). Highlights include:

Shields, R. (2001) The emperor's new clothes. *Trends Genet.* 17, 189

Ashman, K. (2001) Life is sometimes sweet! T cell activation and glycosylation. *Trends Biotechnol.* 16, 161

Woolhouse, M.E.J. (2001) The human genome: what's in it for parasitologists? *Trends Parasitol.* 17, 214

P-val FPScan

Scan **PRINTS** with a PROTEIN query sequence; using an ID code from one of the following databases: {SWISSPROT SPTREMBL SWISSNEW TREMBLNEW} or by pasting it in as a raw sequence.
Please Note; DNA Sequences are NOT catered for in this software.

Important information concerning the E-value calculation [please read](#)

Please input; either an ID code, or a raw sequence:

```
MGFNLTLAKLPNNELHGQESHNSGNRSDGPGKNTTLHNEF
DTIVLPVLYLIIFVASILLNGLAVWIFFHIRNKTSFIFY
KNIVVADLIMTLTFPFRIVHDAGFGPWYFKFILCRYTSV
FYANMYTSIVFLGLISIDRYLKVVKPFGDSRMYSITFTK
LSVCVWVIMAVLSLPNIILTNGOPTEDNIHDCSKLKSPI
VKWHTAVTYVNSCLFVAVLVILIGCYIAISRYIHKSSRC
ISQSSRKRKHNQSIRVVVAVYFTCFLPYHLCRMPSTFSH
```

The E-value threshold determines the level of significance of results in the 1st table

E-value threshold:

Select Database

- ☒ Prints32_0 ☐ Prints30_0 ☐ Blocksplus11
☐ Prints31_0 ☐ Blocks11

Select Matrix

- ☒ blos62
☐ blos45
☐ blos80

Distance variance:
 %

Mail any comments, bugs, or suggestions to:
scordis@bioinf.man.ac.uk

PRINTS32_0 and matrix blos62

Scan of sequence: USER_SEQUENCE

Highest scoring fingerprints for your query

Fingerprint	E-value	GRAPHScan
<u>GPCRRHODOPSN (relations)</u>	3.118054e-29	<u>Graphic</u>

for further information choose any of the following options

- Simple - Top Ten
- Detailed - Top Ten (detailed by motif)

Back to top

Ten top scoring fingerprints for your query

Fingerprint	No. of Motifs	SumId	AveId	PfScore	Pvalue	Evalue	GRAPHScan	
<u>GPCRRHODOPSN</u>	7 of 7	1.8e+02	25	1733	1.2e-34	3.1e-29	iiIiiii	<u>Graphic</u>
<u>PROTEASEAR</u>	2 of 5	58.16	29.08	460	5.2e-08	0.013	i...i.	<u>Graphic</u>
<u>CXCCHMKINER4</u>	2 of 9	79.69	39.84	696	1.5e-07	0.038	.I.I.....	<u>Graphic</u>
<u>DUFFYANTIGEN</u>	3 of 7	61.06	20.35	626	1.8e-06	0.47	i.i...i	<u>Graphic</u>
<u>BRADYKININR</u>	2 of 6	59.29	29.64	419	4.3e-06	1.1	.Ii...	<u>Graphic</u>
<u>P2Y12PRNCPTR</u>	2 of 3	54.53	27.26	466	1.4e-05	3.6	Ii.	<u>Graphic</u>
<u>PAFRECEPTOR</u>	3 of 11	78.40	26.13	677	1.5e-05	3.9	...i...I.i.	<u>Graphic</u>
<u>ANGIOTENSINR</u>	2 of 8	103.09	51.54	435	2.8e-05	7.3	I..I....	<u>Graphic</u>

<u>CELLSNTHASEA</u>	2 of 9	39.25	19.62	445	5.2e-05	13i...i	<u>Graphic</u>
<u>ACRIFLAVINRP</u>	2 of 9	34.38	17.19	318	5.4e-05	14	..i...i..	<u>Graphic</u>

[Back to top](#)

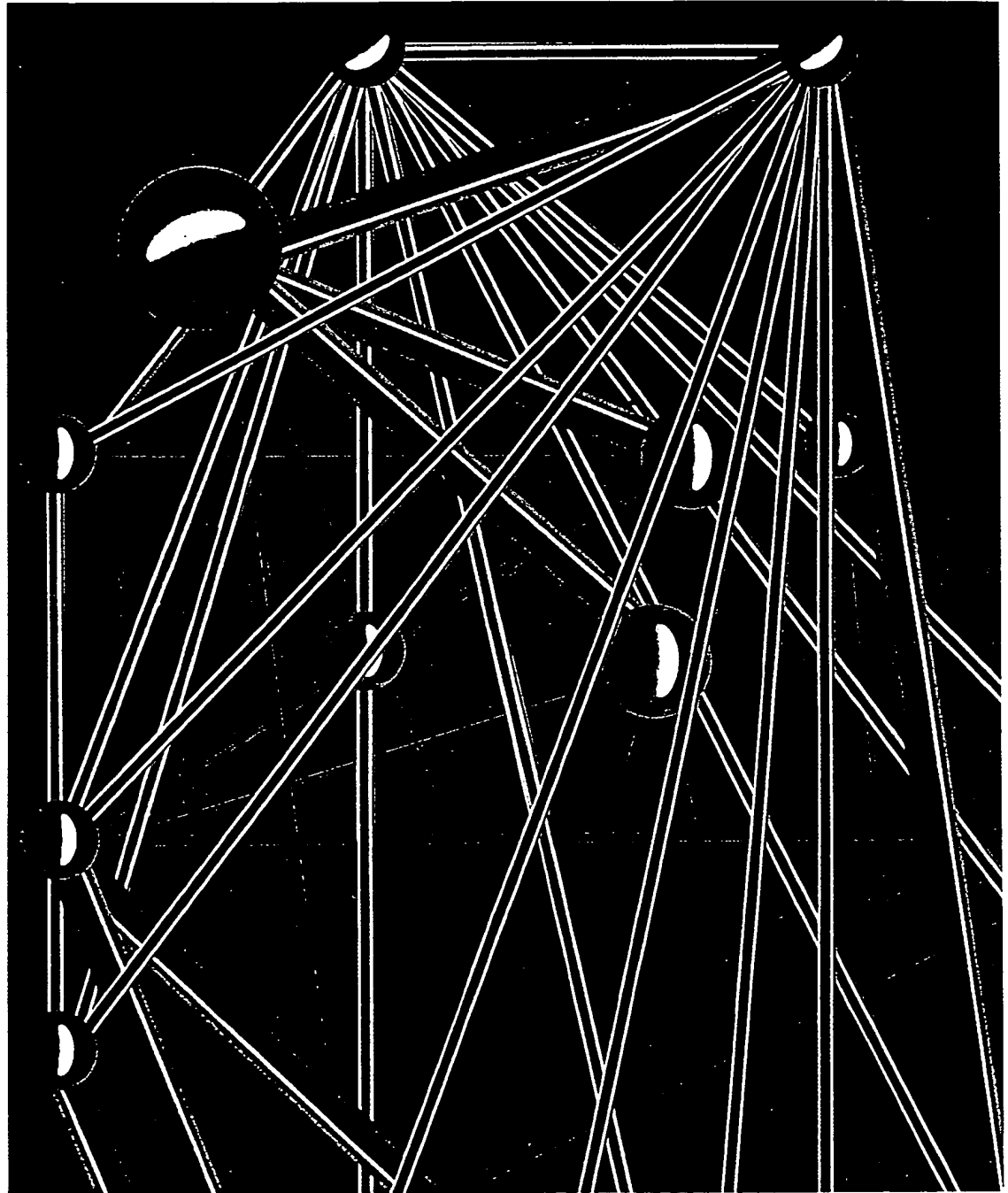
Ten top scoring fingerprints for your query. Detailed by motif					
FingerPrint Name	Motif Number	IdScore	PfScore	Pval	Sequence
GPCRRHODOPSN	1 of 7	23.40	225	8.79e-07	VLPVLYLIIFVASILLNGLAVWIFF
	2 of 7	24.21	190	8.19e-05	FIFYLKNIVVADLIMTLTFPFR
	3 of 7	35.51	339	1.10e-08	FYANMYTSIVFLGLISIDRYLKV
	4 of 7	23.63	251	3.97e-04	FTKVLSVCVWVIMAVLSLPNII
	5 of 7	21.00	134	9.54e-03	VTYVNSCLFVAVLVILIGCYIAIS
	6 of 7	21.06	264	2.05e-04	HNQSIRVVVAVYFTCF
	7 of 7	27.45	330	1.97e-07	KEITLFLSACNVCLDPIIYFFMCRSFS
PROTEASEAR	1 of 5	28.57	236	3.07e-04	KNTTLHNEFDTIVLPVLY
	4 of 5	29.59	224	1.69e-04	QSIRVVVAVYFTCF
CXCCHMKINER4	2 of 9	43.75	346	1.31e-04	HNEFDTIVLPVLYLII
	4 of 9	35.94	350	1.12e-03	GPWYFKFILCRYTSVL
DUFFYANTIGEN	1 of 7	14.88	146	7.40e-02	LPVLYLIIFVASILLNGLAVWIFF
	3 of 7	20.78	263	2.81e-03	ILCRYTSVLFYANMYTSIVFLG
	7 of 7	25.40	217	8.79e-03	HLDRLLESAQKILYYCKEITLFLSAC

BRADYKININR	2 of 6	33.57	245	5.18e-04	YTSVLFYANMYTSI
	3 of 6	25.71	174	8.35e-03	AVLSLPNIILTNGQ
P2Y12PRNCPTR	1 of 3	31.25	208	7.77e-03	SRMYSITFTKVLSVCV
	2 of 3	23.28	258	1.78e-03	HKSSRQFISQSSRKRKHNQSIRVVVAVYF
PAFRECEPTOR	4 of 11	21.15	165	8.94e-02	GLISIDRYLKVVVKPFGDSRMYSITFT
	8 of 11	33.33	332	1.92e-03	PYHLCRMPSTFSLD
	10 of 11	23.91	180	8.91e-02	FFMCRSFSRWLFKKSNIIRPSES
ANGIOTENSINR	1 of 8	60.49	253	1.67e-03	LYLIIFVAS
	4 of 8	42.59	182	1.70e-02	VLSLPNIILTNG
CELLSNTHASEA	5 of 9	15.87	165	5.67e-02	HIRNKTSFIFYLKNIVVADLIMTLTFP
	9 of 9	23.38	280	9.09e-04	TYVNSCLFVAVLVILIGCYIAI
ACRIFLAVINRP	3 of 9	17.92	168	4.03e-03	GKNTTLHNEFDTIVLPVLYLIIFV
	7 of 9	16.46	150	1.34e-02	ITFTKVLSVCVWVIMAVLSLPNII

> USER SEQUENCE

MGFNLTAKLPNNELHGQESHNSGNRSDGPGKNTTLHNEF
 DTIVLPVLYLIIFVASILLNGLAVWIFFHIRNKTSFIFYL
 KNIVVADLIMTLTFPRIVHDAGFGPWYFKFILCRYTSVL
 FYANMYTSIVFLGLISIDRYLKVVVKPFGDSRMYSITFTKV
 LSVCVWVIMAVLSLPNIILTNGQPTEDNIHDCSKLKSPLG
 VKWHTAVTYVNSCLFVAVLVILIGCYIAISRYIHKSSRQF
 ISQSSRKRKHNQSIRVVVAVYFTCFYPYHLCRMPSTFSLD
 DRLLDESAQKILYYCKEITLFLSACNVCLDPIIYFFMCRS
 FSRWLFKKSNIIRPSESIRSLQSVRRSEVRIYYDYTDV

Bioinformatics



David W. Mount

COLD SPRING HARBOR LABORATORY PRESS

BEST AVAILABLE COPY



Bioinformatics

Sequence and Genome Analysis

All rights reserved

© 2001 by Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York

Printed in the United States of America

Developmental Editor Judy Cuddihy

Project Coordinator Joan Ebert

Production Editor Patricia Barker

Interior Designer Denise Weiss

Cover Designer Ed Atkeson, Berg Design

Library of Congress Cataloging-in-Publication Data

Mount, David W.

Bioinformatics : sequence and genome analysis / David W. Mount.

p. cm.

Includes bibliographical references and index.

ISBN 0-87969-597-8 (hard cover : alk. paper)—ISBN 0-87969-608-7 (paperback : alk. paper)

1. Genetics—Data processing. 2. Bioinformatics. 3. Nucleotide sequence. 4. Amino acid sequence. I. Title.

QH441.2 .M68 2000

572.8'633—dc21

00-060252

10 9 8 7 6 5 4 3 2 1

Front cover: Illustration inspired by the relationship of the 6000 genes in yeast to each other (Lisa Mount and Adam Sherman).

Students and researchers using the procedures in this manual do so at their own risk. Cold Spring Harbor Laboratory makes no representations or warranties with respect to the material set forth in this manual and has no liability in connection with the use of these materials.

Laser radiation, visible or invisible, can cause severe damage to the eyes and skin. Take proper precautions to prevent exposure to direct and reflected beams. Always follow manufacturers' safety guidelines and consult your local safety office.

Procedures for the humane treatment of animals must be observed at all times. Check with the local animal facility for guidelines.

All WorldWideWeb addresses are accurate to the best of our knowledge at the time of printing.

Certain experimental procedures in this manual may be the subject of national or local legislation or agency restrictions. Users of this manual are responsible for obtaining the relevant permissions, certificates, or licenses in these cases. Neither the authors of this manual nor Cold Spring Harbor Laboratory assumes any responsibility for failure of a user to do so.

The polymerase chain reaction process is covered by certain patent and proprietary rights. Users of this manual are responsible for obtaining any licenses necessary to practice PCR or to commercialize the results of such use. COLD SPRING HARBOR LABORATORY MAKES NO REPRESENTATION THAT USE OF THE INFORMATION IN THIS MANUAL WILL NOT INFRINGE ANY PATENT OR OTHER PROPRIETARY RIGHT.

Authorization to photocopy items for internal or personal use, or the internal or personal use of specific clients, is granted by Cold Spring Harbor Laboratory Press, provided that the appropriate fee is paid directly to the Copyright Clearance Center (CCC). Write or call CCC at 222 Rosewood Drive, Danvers, MA 01923 (508-750-8400) for information about fees and regulations. Prior to photocopying items for educational classroom use, contact CCC at the above address. Additional information on CCC can be obtained at CCC Online at <http://www.copyright.com/>

All Cold Spring Harbor Laboratory Press publications may be ordered directly from Cold Spring Harbor Laboratory Press, 10 Skyline Drive, Plainview, New York 11803-2500. Phone: 1-800-843-4388 in Continental U.S. and Canada. All other locations: (516) 349-1930. FAX: (516) 349-1946. E-mail: cshpress@cshl.org. For a complete catalog of all Cold Spring Harbor Laboratory Press publications, visit our World Wide Web site at <http://www.cshl.org/>

(back cover)

Bioinformatics

SEQUENCE AND GENOME ANALYSIS

The application of computational methods to DNA and protein science is a new and exciting development in biology. *Bioinformatics: Sequence and Genome Analysis* is a comprehensive introduction to this emerging field of study. The book has many unique and valuable features:

Essential for any biologist who wants to understand methods of sequence and structure analysis and how the necessary computer programs work.

Sequence alignment, structure prediction, phylogenetic and gene prediction, database searching, and genome analysis are clearly explained and amply illustrated.

Underlying algorithms and assumptions are clearly explained for the non-specialist.

Examples are presented in simple numerical terms rather than complex formulas and notation.

Theoretical underpinnings are linked to biological problems and their solutions.

Extensive tables provide descriptions and Web sources for a broad range of publicly available software.

An associated Website (www.bioinformaticsonline.org), accessible free of charge by book purchasers, provides links to Internet sources referred to in the text, as well as problem sets for classroom use, and other useful material not included in the text.

Based on a well-established course given at the University of Arizona by the author, David Mount, this book is an ideal foundation for teaching at an undergraduate and graduate level. It is also highly suited for the self-instruction of investigators interested in the application of methods and strategies in functional genomics and for the needs of information specialists working in molecular biology and pharmaceutical laboratories.

www.bioinformaticsonline.org

ISBN 0-87969-597-8



INTRODUCTION

DATABASE SIMILARITY SEARCHES have become a mainstay of bioinformatics. Large sequencing projects in which all the genomic DNA sequence of an organism is obtained have become quite commonplace. The genomes of a number of model organisms have been sequenced, including the budding yeast *Saccharomyces cerevisiae*, the bacterium *Escherichia coli*, the worm *Caenorhabditis elegans*, the fruit fly *Drosophila melanogaster*, and the human species *Homo sapiens*. These species have also been subjected to intense biological analysis to discover the functions of the genes and encoded proteins. Thus, there is a good deal of information available as to the biological function of particular sequences in model organisms that may be exploited to predict the function of similar genes in other organisms. In addition to genomic DNA sequences, complete cDNA copies of messenger RNAs that carry all the sequence information for the protein products have also been obtained for some of the expressed genes of various organisms. Translation of these cDNA copies provides a close-to-correct prediction of the sequence of the encoded proteins. Because obtaining intact cDNA sequences is laborious and time-consuming, a common practice is to make a library of partial cDNA sequences from the expressed genes, and then to perform high-throughput, low-accuracy sequencing of a large number of these partial sequences known as expressed sequence tags (ESTs). The objective of an EST project is to find enough sequence of each cDNA and to have enough accuracy in the sequence that the amino acid sequence of a significant length of the encoded protein can be predicted. Overlapping ESTs can then be combined, and interesting ones can be found by database similarity searches. The full cDNA sequence of these genes of interest may then be obtained. Once all the sequence information is collected and placed in the sequence databases, the big task at hand is to search through the databases to locate similar sequences that are predicted to have a similar biological function through a close evolutionary relationship.

Sequence database searches can also be remarkably useful for finding the function of genes whose sequences have been determined in the laboratory. The sequence of the gene of interest is compared to every sequence in a sequence database, and the similar ones are identified. Alignments with the best-matching sequences are shown and scored. If a query sequence can be readily aligned to a database sequence of known function, structure, or biochemical activity, the query sequence is predicted to have the same function, structure, or biochemical activity. The strength of these predictions depends on the quality of the alignment between the sequences. As a rough rule, if more than one-half of the amino acid sequence of query and database proteins is identical in the sequence alignments, the prediction is very strong. As the degree of similarity decreases, confidence in the prediction also decreases. The programs used for these database searches provide statistical evaluations that serve as a guide for evaluation of the alignment scores.

Previous chapters have described methods for aligning sequences or for finding common patterns within sequences. The purpose of making alignments is to discover whether or not sequences are homologous or derived from a common ancestor gene. If a homology relationship can be established, the sequences are likely to have maintained the same function as they diverged from each other during evolution. If an alignment can be found that would rarely be observed between random sequences, the sequences are predicted to be related with a high degree of confidence. The presence of one or more conserved patterns in a group of sequence is also useful for establishing evolutionary and structure-function relationships among sequences.

The above methods of establishing sequence relationships have been utilized in database searches that are summarized in Table 7.1. In addition to standard searches of a sequence

Algorithms on Strings, Trees, and Sequences

**COMPUTER SCIENCE AND COMPUTATIONAL
BIOLOGY**

Dan Gusfield

University of California, Davis



**CAMBRIDGE
UNIVERSITY PRESS**

BEST AVAILABLE COPY

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge CB2 1RP
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, Melbourne 3166, Australia

© Dan Gusfield 1997

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 1997
Reprinted 1999 (with corrections)

Printed in the United States of America

Typeset in Times Roman

Library of Congress Cataloging-in-Publication Data

Gusfield, Dan.

Algorithms on strings, trees, and sequences : computer science and
computational biology / Dan Gusfield.1

p. cm.

ISBN 0-521-58519-8 (hc)

1. Computer algorithms. 2. Molecular biology – Data processing.

I. Title.

QA76.9.A43387 1997

005.7'3 – dc21

86-46612

A catalog record for this book is available from the British Library.

ISBN 0 521 58519 8 hardback

BEST AVAILABLE COPY

The Importance of (Sub)sequence Comparison in Molecular Biology

Sequence comparison, particularly when combined with the systematic collection, curation, and search of databases containing biomolecular sequences, has become essential in modern molecular biology. Commenting on the (then) near-completion of the effort to sequence the entire yeast genome (now finished), Stephen Oliver says

In a short time it will be hard to realize how we managed without the sequence data. Biology will never be the same again. [478]

One fact explains the importance of molecular sequence data and sequence comparison in biology.

The first fact of biological sequence analysis

The first fact of biological sequence analysis In biomolecular sequences (DNA, RNA, or amino acid sequences), high sequence similarity usually implies significant functional or structural similarity.

Evolution reuses, builds on, duplicates, and modifies “successful” structures (proteins, exons, DNA regulatory sequences, morphological features, enzymatic pathways, etc.). Life is based on a repertoire of structured and interrelated molecular building blocks that are shared and passed around. The same and related molecular structures and mechanisms show up repeatedly in the genome of a single species and across a very wide spectrum of divergent species. “Duplication with modification” [127, 128, 129, 130] is the central paradigm of protein evolution, wherein new proteins and/or new biological functions are fashioned from earlier ones. Doolittle emphasizes this point as follows:

The vast majority of extant proteins are the result of a continuous series of genetic duplications and subsequent modifications. As a result, redundancy is a built-in characteristic of protein sequences, and we should not be surprised that so many new sequences resemble already known sequences. [129]

He adds that

... all of biology is based on an enormous redundancy ... [130]

The following quotes reinforce this view and suggest the utility of the “enormous redundancy” in the practice of molecular biology. The first quote is from Eric Wieschaus, cowinner of the 1995 Nobel prize in medicine for work on the genetics of *Drosophila* development. The quote is taken from an Associated Press article of October 9, 1995. Describing the work done years earlier, Wieschaus says

We didn't know it at the time, but we found out everything in life is so similar, that the same genes that work in flies are the ones that work in humans.

And fruit flies aren't special. The following is from a book review on DNA repair [424]:

Throughout the present work we see the insights gained through our ability to look for sequence homologies by comparison of the DNA of different species. Studies on yeast are remarkable predictors of the human system!

So "redundancy", and "similarity" are central phenomena in biology. But similarity has its limits – humans and flies do differ in some respects. These differences make *conserved* similarities even more significant, which in turn makes *comparison* and *analogy* very powerful tools in biology. Lesk [297] writes:

It is characteristic of biological systems that objects that we observe to have a certain form arose by evolution from related objects with similar but not identical form. They must, therefore, be robust, in that they retain the freedom to tolerate some variation. We can take advantage of this robustness in our analysis: By identifying and comparing related objects, we can distinguish variable and conserved features, and thereby determine what is crucial to structure and function.

The important "related objects" to compare include much more than sequence data, because biological universality occurs at many levels of detail. However, it is usually easier to acquire and examine sequences than it is to examine fine details of genetics or cellular biochemistry or morphology. For example, there are vastly more protein sequences known (deduced from underlying DNA sequences) than there are known three-dimensional protein structures. And it isn't just a matter of convenience that makes sequences important. Rather, the biological sequences *encode* and reflect the more complex common molecular structures and mechanisms that appear as features at the cellular or biochemical levels. Moreover, "nowhere in the biological world is the Darwinian notion of 'descent with modification' more apparent than in the sequences of genes and gene products" [130]. Hence a tractable, though partly heuristic, way to search for functional or structural universality in biological systems is to search for similarity and conservation at the *sequence* level. The power of this approach is made clear in the following quotes:

Today, the most powerful method for inferring the biological function of a gene (or the protein that it encodes) is by sequence similarity searching on protein and DNA sequence databases. With the development of rapid methods for sequence comparison, both with heuristic algorithms and powerful parallel computers, discoveries based solely on sequence homology have become routine. [360]

Determining function for a sequence is a matter of tremendous complexity, requiring biological experiments of the highest order of creativity. Nevertheless, with only DNA sequence it is possible to execute a computer-based algorithm comparing the sequence to a database of previously characterized genes. In about 50% of the cases, such a mechanical comparison will indicate a sufficient degree of similarity to suggest a putative enzymatic or structural function that might be possessed by the unknown gene. [91]

Thus large-scale sequence comparison, usually organized as database search, is a very powerful tool for biological inference in modern molecular biology. And that tool is almost universally used by molecular biologists. It is now standard practice, whenever a new gene is cloned and sequenced, to translate its DNA sequence into an amino acid sequence and then search for similarities between it and members of the protein databases. No one today would even think of publishing the sequence of a newly cloned gene without doing such database searches.

The final quote reflects the potential total impact on biology of the *first fact* and its exploitation in the form of sequence database searching. It is from an article [179] by Walter Gilbert, Nobel prize winner for the coinvention of a practical DNA sequencing method. Gilbert writes:

The new paradigm now emerging, is that all the 'genes' will be known (in the sense of being resident in databases available electronically), and that the starting point of biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis.

Already, hundreds (if not thousands) of journal publications appear each year that report biological research where sequence comparison and/or database search is an integral part of the work. Many such examples that support and illustrate the *first fact* are distributed throughout the book. In particular, several in-depth examples are concentrated in Chapters 14 and 15 where multiple string comparison and database search are discussed. But before discussing those examples, we must first develop, in the next several chapters, the techniques used for approximate matching and (sub)sequence comparison.

Caveat

The *first fact of biological sequence analysis* is extremely powerful, and its importance will be further illustrated throughout the book. However, there is not a one-to-one correspondence between sequence and structure or sequence and function, because the converse of the *first fact* is not true. That is, high sequence similarity usually implies significant structural or functional similarity (the first fact), but structural or functional similarity does not necessarily imply sequence similarity. On the topic of protein structure, F. Cohen [106] writes "... similar sequences yield similar structures, but quite distinct sequences can produce remarkably similar structures". This *converse* issue is discussed in greater depth in Chapter 14, which focuses on multiple sequence comparison.

ISI HOME | ABOUT ISI | PRODUCTS | SUPPORT | JOURNAL LISTS | CONTACT US | EMPLOY
SEARCH | WHAT'S NEW IN RESEARCH | NEWS | LANGUAGES | PRIVACY

Most-Cited Papers of 1990-98

1990:

S.F. Altschul, *et al.*, "Basic Local Alignment Search Tool," *J. Mol. Biol.*, 215:403, 1990. **Citations: 9,969**

1991:

S. Moncada, R.M.J Palmer, E.A. Higgs, "Nitric Oxide: Physiology, pathophysiology, and pharmacology," *Pharm. Rev.*, 43:109, 1991. **Citations: 6,655**

1992:

R.O. Hynes, "Integrins: Versatility, modulation, and signaling in cell adhesion," *Cell*, 69:11, 1992. **Citations: 4,610**

1993:

M.J. Berridge, "Inositol trisphosphate and calcium signaling," *Nature*, 361:315, 1993. **Citations: 3,446**

1994:

J.D. Thompson, D.G. Higgins, T.J. Gibson, "Clustal W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice," *Nucl. Acids Res.*, 22:4673, 1994. **Citations: 3,352**

1995:

C.B. Thompson, "Apoptosis in the pathogenesis and treatment of disease," *Science*, 267:1456, 1995. **Citations: 1,745**

1996:

R.M. Barnett, *et al.*, "Particles and fields. 1. Review of particle physics," *Phys. Rev. D*, 54:1, 1996. **Citations: 1,342**

1997:

S.F. Altschul, *et al.*, "Gapped BLAST and Psi-BLAST: A new generation of protein database search programs," *Nucl. Acids Res.*, 25:3389, 1997. **Citations: 1,534**

1998:

S.H. Landis, *et al.*, "Cancer statistics, 1998," *CA-A Canc. J.*, 48:6, 1998. **Citations: 605**

Source: [High-Impact Papers](#)



webmaster@isinet.com

Expression cloning of a cDNA encoding the bovine histamine H₁ receptor

(adrenal medulla/*Xenopus* oocyte/[³H]mepyramine/doxepin)

MASAKATSU YAMASHITA*, HIROYUKI FUKUI*†, KAZUSHIGE SUGAMA‡, YOSHIYUKI HORIO*, SEIJI ITO‡, HIROYUKI MIZUGUCHI*, AND HIROSHI WADA*

*Department of Pharmacology II, Faculty of Medicine, Osaka University, Suita 565, Japan; and †Department of Cell Biology, Osaka Bioscience Institute, Suita 565, Japan

Communicated by Esmond E. Snell, September 13, 1991

ABSTRACT A functional cDNA clone for the histamine H₁ receptor was isolated from a cDNA library of bovine adrenal medulla by a combination of molecular cloning in an expression vector and electrophysiological assay in *Xenopus* oocytes. The H₁ receptor cDNA encodes a protein of 491 amino acids (*M*_r 55,954) with seven putative transmembrane domains, illustrating the similarity to other receptors that couple with guanine nucleotide-binding regulatory proteins (G protein-coupled receptors). The sequence homology between the H₁ and H₂ receptors is not higher than that between the histamine H₁ and m₁-muscarinic receptors. The cloned receptor protein expressed in COS-7 cells bound specifically to [³H]mepyramine, an H₁ receptor antagonist, and this binding was displaced by H₁ receptor antagonists and histamine with affinities comparable with those in membranes of bovine adrenal medulla. H₁ receptor mRNA was shown to be expressed in brain and in peripheral tissues, including lung, small intestine, and adrenal medulla. This investigation discloses the molecular nature of the H₁ receptor—a receptor that mediates diverse neuronal and peripheral actions of histamine and that may be of therapeutic importance in allergy.

Since Dale and Laidlaw (1) first reported the contraction of smooth muscle by histamine, the pharmacological significance of this phenomenon has been extensively investigated. Three subtypes of histamine receptor (H₁, H₂, and H₃) are known. The H₁ receptor was identified by Ash and Schild (2) and H₁ receptor antagonists have been used in the therapy of many allergic diseases, including urticaria, allergic rhinitis, pollenosis, and bronchial asthma. In peripheral tissues, the histamine H₁ receptor mediates the contraction of smooth muscles, increase in capillary permeability due to contraction of terminal venules, and catecholamine release from adrenal medulla (3), as well as mediating neurotransmission in the central nervous system (4). Although signal transduction of the H₁ receptor through Ca²⁺ mobilization via an increase in the intracellular inositol 1,4,5-trisphosphate level has been extensively investigated (5, 6), little is known about the molecular structure of the histamine H₁ receptor. Recently, another method for cDNA cloning of Ca²⁺-mobilizing receptors through their expression in *Xenopus* oocytes has been developed (7). Meyerhof *et al.* (8) and Sugama *et al.* (9) have reported that the injection of poly(A)⁺ RNA prepared from bovine adrenal medulla into *Xenopus* oocytes resulted in functional expression of the histamine H₁ receptor in oocytes. The present study describes the cloning and sequencing of a cDNA encoding histamine H₁ receptor³ from a cDNA library of bovine adrenal medulla using *in vitro* RNA

transcription and electrophysiological assay with *Xenopus* oocytes.

MATERIALS AND METHODS

Materials. [³H]Mepyramine (1073 GBq/mmol) and [α -³²P]-dCTP (\approx 111 TBq/mmol) were purchased from DuPont/NEN. Histamine and (+)-chlorpheniramine were purchased from Wako Pure Chemical (Osaka) and Tokyo Kasei (Tokyo), respectively. Mepyramine and doxepin were purchased from Sigma. (–)-Chlorpheniramine and famotidine were gifts from Smith Kline & French and Yamanouchi Pharmaceutical (Tokyo), respectively. A mammalian expression vector pEF-BOS (10) was donated by S. Nagata of the Osaka Bioscience Institute.

Isolation of Poly(A)⁺ RNA. Total RNA was extracted by the acid guanidinium isothiocyanate/phenol/chloroform method (11). Poly(A)⁺ RNA was isolated by chromatography on oligo(dT)-cellulose (12).

Expression Cloning of Histamine H₁ Receptor cDNA. Bovine adrenal medullary poly(A)⁺ RNA (\approx 180 μ g) was size-fractionated on a 5–25% (wt/vol) sucrose-density gradient. An aliquot (1 μ l) of each poly(A)⁺ RNA fraction (20 μ l) was injected into *Xenopus* oocytes, and electrophysiological assay by measuring Ca²⁺-dependent inward Cl[–] currents was done as described (9). The fraction that showed the highest histamine-induced inward Cl[–] currents was used for oligo(dT)-primed cDNA synthesis. Double-stranded cDNAs of >2-kilobase (kb) pairs were size-selected by agarose gel electrophoresis followed by elution with GeneClean II (Bio 101, La Jolla, CA) and were ligated into λ ZAPII (Stratagene) at the *Eco*RI site. The library was divided and amplified in 65 pools of \approx 20,000 independent clones each. *In vitro* transcription was done essentially according to the procedure of Julius *et al.* (13). RNA transcripts (\approx 5 ng) from each pool were individually injected into *Xenopus* oocytes. After incubation for 1–2 days, the oocytes were tested for inward Cl[–] currents induced by 100 μ M histamine under a voltage-clamp at –60 mV. The single positive pool of 20,000 clones was progressively subdivided into smaller pools of 8000, 4000, 400, and 15 clones until finally a single clone was obtained. cDNA encoding the histamine H₁ receptor was sequenced by the M13 chain-termination method (14) using a DNA sequencer (model 370A, Applied Biosystems). The sequence homology search was done by using DNASIS (Hitachi Software Engineering, Yokohama, Japan).

Expression of Histamine H₁ Receptor in COS-7 Cells and Its Determination by [³H]Mepyramine-Binding Assay. An *Eco*RI

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

[†]To whom reprint requests should be addressed at: Department of Pharmacology II, Faculty of Medicine, Osaka University, 2-2 Yamadaoka, Suita 565, Japan.

³The sequence reported in this paper has been deposited in the GenBank data base (accession no. D90430).

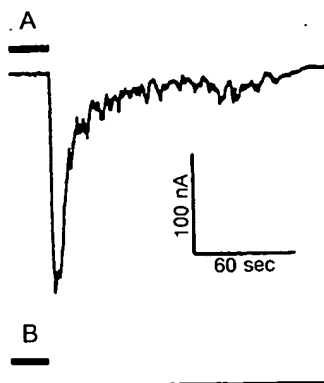


FIG. 1. (A) Current trace recorded from a *Xenopus* oocyte injected with *in vitro* synthesized histamine H₁ receptor mRNA. (B) Mepyramine (10 μ M) was administered 30 sec before histamine application. Recordings were obtained at a voltage-clamped membrane potential of -60 mV. Concentration of histamine applied was 100 μ M; horizontal bar indicates duration of application. Data were reproducible ($n = 5$), and representative tracings are shown.

fragment (2.7 kb) of the H₁ receptor cDNA was subcloned into the mammalian expression vector pEF-BOS at the *Bsr*XI site. COS-7 cells were transfected by the DEAE-dextran method and were harvested after 60 hr (15). Preparation of membranes from COS-7 cells and [³H]mepyramine-binding assay were done by a described method (16). Nonspecific bindings of [³H]mepyramine to both transfected and nontransfected cells at 2.6 nM radioligand were <10% of total binding to nontransfected cells. Specific binding of [³H]mepyramine to the nontransfected cells was observed (basal control), but that from the transfected cells assayed with 2.6 nM [³H]mepyramine (3.4 pmol/mg of protein) was ≈30 times the basal control (0.1 pmol/mg of protein). Specific binding of [³H]mepyramine to the expressed binding site was calculated by subtracting specific [³H]mepyramine binding to the nontransfected cells from that to the transfected cells.

RNA Blot Analysis. Poly(A)⁺ RNA prepared from various bovine tissues was separated (7 μ g per lane) by formaldehyde/1% agarose gel electrophoresis (17) and transferred to a nylon membrane (Schleicher & Schuell). A 2.7-kb *Eco*RI fragment of the histamine H₁ receptor cDNA was labeled with [α -³²P]dCTP by the random-priming method and was

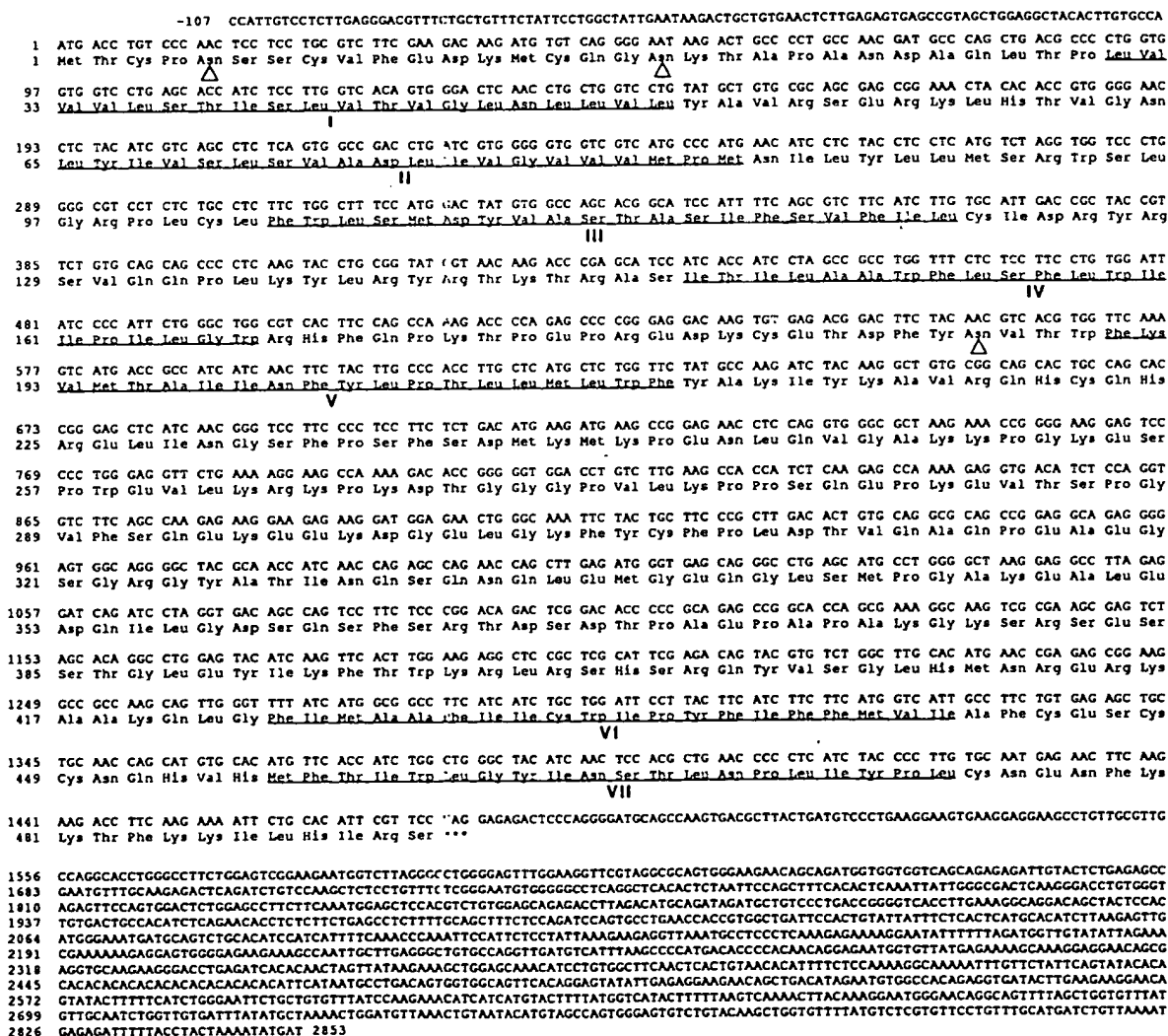


Fig. 2. Nucleotide and deduced amino acid sequences of the histamine H₁ receptor cDNA clone. Sequences of both strands of cDNA were determined. Positions of the putative transmembrane segments I–VII of the H₁ receptor are indicated below amino acid sequence; the terminal of each segment is tentatively assigned from a hydropathy profile. Triangles indicate potential N-glycosylation sites.

used as a probe (18). Hybridization was done at 42°C in 5× standard saline citrate/20 mM sodium phosphate, pH 7.0/1× Denhardt's solution/50% (vol/vol) formamide/0.1% SDS/10% (wt/vol) dextran sulfate/salmon sperm DNA at 100 µg/ml. The membrane was washed with 0.1× standard saline citrate and 0.1% SDS at 42°C.

RESULTS

Isolation of a Histamine H₁ Receptor cDNA. Poly(A)⁺ RNA isolated from bovine adrenal medulla was size-fractionated in a sucrose-density gradient. Two peaks giving histamine-evoked inward currents in oocytes were observed in the size range of 2.5- to 3.5-kb nucleotides and above 5-kb nucleotides (data not shown). A cDNA library was constructed from poly(A)⁺ RNA in the fraction of 2.5- to 3.5-kb nucleotides giving the highest response. Of 65 pools tested only one pool gave small inward currents in response to 100 µM histamine. After several subdivisions of the positive pool, a single clone encoding for a functional histamine H₁ receptor was isolated; histamine induced inward Cl⁻ currents in oocytes injected with *in vitro*-transcribed mRNA from the cloned histamine H₁ receptor cDNA (Fig. 1), and mepyramine, an H₁ receptor antagonist, at 10⁻⁶ M completely blocked the histamine-induced response in oocytes.

Primary Structure of the Histamine H₁ Receptor. The nucleotide and deduced amino acid sequences of the bovine histamine H₁ receptor are shown in Fig. 2. The clone (2960 nucleotides long) consisted of 107 nucleotides of the 5' untranslated region, 1473 nucleotides of the coding region, and 1380 nucleotides of the 3'-untranslated region. The histamine H₁ receptor cDNA encodes a protein of 491 amino acids with a M_r of 55,954.

Pharmacological Characterization of [³H]Mepyramine-Binding to the Histamine H₁ Receptor Expressed in COS-7 Cells. For determination of pharmacological characters of the receptor, the *Eco*RI fragment (2.7 kb) of the H₁ receptor cDNA was subcloned into the mammalian expression vector pEF-BOS, and the vector was introduced into monkey kidney COS-7 cells. After 60-hr incubation, the binding of [³H]mepyramine to the membranes from the cells was measured. Specific binding of [³H]mepyramine to the expressed binding site was saturable, and Scatchard plot analysis indicated the presence of a single binding site with a K_d value of 3.2 nM and a B_{max} value of 6.6 pmol/mg of protein (Fig. 3A). K_i values of mepyramine, and (+)- and (-)-chlorphenir-

amines were determined to be 2.6 × 10⁻⁹ M, 8.0 × 10⁻⁹ M, and 7.6 × 10⁻⁷ M, respectively (Fig. 3B). These K_d and K_i values and the stereoselectivity of (+)- and (-)-chlorpheniramines for the binding site expressed in COS-7 cells were comparable with those for adrenal medullary membranes. The K_d value was 1.5 × 10⁻⁹ M; K_i values were 1.8 × 10⁻⁹ M (mepyramine), 4.3 × 10⁻⁹ M [(+)-chlorpheniramine], and 4.6 × 10⁻⁷ M [(-)-chlorpheniramine], as described (19).

Tissue Distribution of Histamine H₁ Receptor mRNA. Tissue distribution of receptor mRNA was determined by RNA blot analysis (Fig. 4). A band of 3.0-kb nucleotides corresponding to a histamine H₁ receptor mRNA was detected in various bovine tissues. The level of H₁ receptor mRNA was high in the lung and small intestine, moderate in the adrenal medulla and uterus, and lower in the cerebral cortex and spleen. No H₁ receptor mRNA was detectable in the cardiac atrium or liver.

DISCUSSION

In the present study, we isolated and sequenced a cDNA clone for the bovine histamine H₁ receptor by using an oocyte expression system and also examined the pharmacological properties of this receptor and the tissue distribution of its mRNA.

The cloned cDNA had no poly(A)⁺, but its size [2960 base pairs (bp)] was comparable with that of histamine H₁ receptor mRNA determined by RNA blot analysis. The M_r of encoded H₁ receptor (55,954) was also consistent with the values estimated by photoaffinity labeling of bovine adrenal medulla (M_r 53,000–58,000) (19) and in guinea pig tissues (M_r 56,000–57,000) (20). Hydropathy-profile analysis (21) of the histamine H₁ receptor revealed the existence of seven putative transmembrane domains, indicating a similar topology to those proposed for other G protein-coupled receptors. The histamine H₁ receptor also possesses a characteristic large third cytoplasmic loop and short carboxyl terminus (22), as do the m₁-muscarinic (23) and dopamine-D₂ (24) receptors. We observed another ATG codon 39 bp downstream from the presumed initiation codon. Comparison with Kozak consensus sequence (25) indicated that neither of the two ATG codons had any advantage as an initiation codon. However, as receptors for biogenic amines and acetylcholine possess conservative aspartate residues at position 108 as putative binding sites for their monoamine and tertiary-amine residues (26), we presume that the upstream ATG codon is the

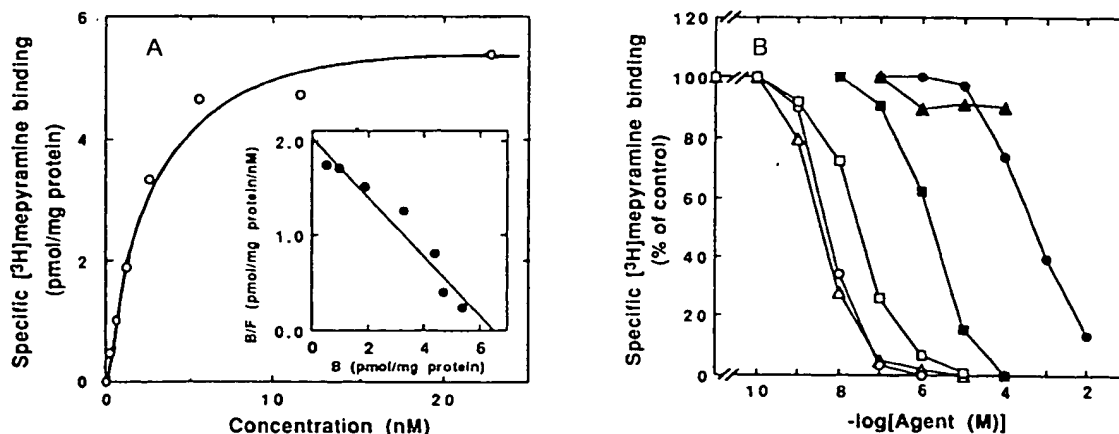


FIG. 3. Binding of [³H]mepyramine to transfected COS-7 cell membranes. (A) Saturation isotherm of specific binding of [³H]mepyramine to membranes from COS-7 cells transfected with the receptor cDNA (O). (Inset) Scatchard plot of this data. B/F, bound/free. (B) Inhibition of [³H]mepyramine-binding to transfected COS-7 cell membranes by various drugs. Membranes were incubated with 4 nM [³H]mepyramine and various concentrations of doxepin (Δ), mepyramine (O), (+)-chlorpheniramine (□), (-)-chlorpheniramine (■), famotidine (▲), or histamine (●). Data points are means of triplicate experiments.

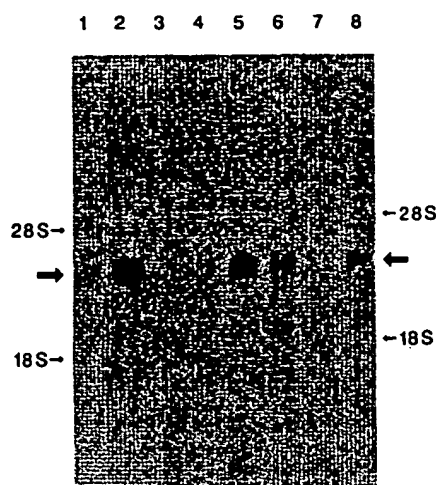


Fig. 4. RNA blot analysis of mRNA isolated from various bovine tissues. Lanes contain 7- μ g samples of poly(A)⁺ RNA from cerebral cortex (lane 1), lung (lane 2), liver (lane 3), cardiac atrium (lane 4), small intestine (lane 5), adrenal medulla (lane 6), spleen (lane 7), and uterus (lane 8). Arrow indicates H₁ receptor mRNA.

initiation codon because it would give a histamine H₁ receptor with the conservative aspartate residue at position 108. The histamine H₁ receptor is highly similar to other G protein-coupled receptors. The sequence of the histamine H₁ receptor is compared with those of some other G protein-coupled receptors in Fig. 5. Sequence homology of transmembrane domains between H₁ and H₂ receptors (40.7%) (27) is not higher than that between H₁ and m₁-muscarinic receptors (44.3%) (23).

There are two potential N-glycosylation sites (Asn-5, Asn-18) in the amino-terminal region with a consensus sequence Asn-Xaa-Ser/Thr (Fig. 2) (29). Mitsuhashi and Payan (30) reported regulation of the affinity of the histamine H₁ receptor by its glycosylation. An additional N-glycosylation site (Asn-187) was observed in the second extracellular loop of the cloned receptor.

The third cytoplasmic loop of the histamine H₁ receptor, which, by analogy, is thought to interact with a G protein, has

many serine and threonine residues that may serve as sites for phosphorylation by protein kinases (Fig. 2). Signal transduction through the histamine H₁ receptor is depressed by activation of protein kinase C in various cells (31–33). Thus, the potential sites of phosphorylation in the third cytoplasmic loop may play an important role in regulating signal transduction through the receptor molecule.

Amino acid residues that are conserved in G protein-coupled receptors were also seen in the H₁ receptor: (i) Two cysteines (Cys-101 and Cys-181) that have been proposed to form a disulfide bond appear in the first and the second extracellular loops (34). (ii) An aspartate residue (Asp-74) is present in the second transmembrane domain. (iii) An anionic and cationic amino acid pair (Asp-125 and Arg-126) occurs at the cytoplasmic border of the third transmembrane domain. (iv) A conservative sequence of 10 amino acids (Leu-460–Pro-469) is observed in the seventh transmembrane domain.

The H₁ receptor mRNA was visualized by RNA blot analysis in various bovine tissues in which the existence of H₁ receptors was reported (3). The presence of the H₁ receptor mRNA in bovine uterus was clearly demonstrated, whereas only H₂ receptors (35) and both H₁ and H₂ receptors (36) were reported present in the uterus from pharmacological studies. The band of H₁ receptor mRNA from brain was unexpectedly faint (Fig. 4); this observation was surprising because the [³H]mepyramine-binding capacities of brain membranes from various species are reported comparable to those of membranes from peripheral tissues (6). Doxepin is a potent displacer of [³H]mepyramine bound to the histamine H₁ receptor from bovine adrenal medulla (Fig. 3). A doxepin-insensitive subtype of histamine H₁ receptor has been proposed to be present in brain because the binding capacity of [³H]doxepin to rat brain membranes is $\approx 10\%$ that of [³H]mepyramine (37).

Cardiac atrium and liver did not give detectable bands of H₁ receptor mRNA (Fig. 4). Pharmacological studies indicate the presence of H₁ receptors in heart (3). However, biochemical results (20) show that the *M_r* of the histamine H₁ receptor in guinea pig heart is 68,000, which is larger than the sizes (*M_r* 56,000–57,000) of these receptors in lung, intestine, and cerebellum, suggesting a subtype of H₁ receptors in heart in which the H₁ receptor mRNA does not hybridize with the cloned cDNA. A relatively large amount of [³H]mepyramine-binding protein is present in liver and was recently suggested

	33	I	II	III
H1	--VLTSTISLVTGGLLVLVAVRSERKLTGVLNIVLSVADLVGVVPHNIIYL--LMS--RWSLGRPLCLFWLSMDYVASTAFISVFLICIDRYRSVQOPLKY--			
H2	--VVLTVLILITAGHVVVCLAVGLNRRSLTNCFLVSLAITDGLLGLVLPFSAFYQ--LSC--RWSFGKRVCFNITYSLDVHCLTASILNLFMSISDRYCAVTDPLRY--			
M1	--STGLLSLATVGLLVLISIKVNTLKTNNYFLSLACADLIIGTFSNLIYTYL--LNG--HWALGTACDLWALDLYVSNAGVHMLLISIDRYSVQPLSY--			
$\alpha 1$	--VILGGLILFVGLGILVLSVACRRRLHSVTHYIIVNLAVADLLTSTVLPFAIFE--ILG--YMAFGVFCVNVAAVDVLCCTASINGLCIISIDRYGVSYPLRY--			
5HT-1c	--LSIVVLIINTIGGILVIMAVSMKKLNATNYFLMSLAIDHVLGLLVNPLSLAI--LYDYVPLPRYLCPVWISLDVLFSTASIMELCAISIDRYVAIRNPIDH--			
D2	--MLTLLIFIIVFGHVLVCHAVSREKALQTTNIVLSLAVADLLVATLVHP--WVYLVGVG--EWKFSRIHCDIFVTLDVHCTASILNLCAISIDRYTAVAMPNLYN			
	137	IV	V	
H1	LRRTKTRASITILAAWFLSF--LWI--IPI--LQWRHFQKTP--EPREDKCTDFYVNTVFKVNTAIIINFTYLP--TLLMLNFKYAKIYKAVRQHCORRELINGSF--(180 a.a.)--			
H2	PVLIPTVRVAVSLVINVISITLSF--LSIHGLWNSRNETSPFNHTIPCKVQV--NLV--YGLVDGLVTFYLP--LLVMCIITYRIFKIAIRRIHQAKHMSWK--(4 a.a.)--			
M1	RAKRTPRRAALHIGLAWLSVFLWA--PAI--LFW--QYLUGER--TVLAGQCYIQFLSQP--IITFGTAAAPYLP--VTVMCTLYWRIYRETNARELAALQGE--(129 a.a.)--			
$\alpha 1$	PTIVTQKRGILLALLCVWALSIVISI--GFL--FCWR--QP--AP--E--DETICQIN--EEPQVFLSALGSFTVP--LTILLVMYCRVYVAKRESGLSKLTD--(54 a.a.)--			
5HT-1c	SRFNSRTKAIKILAIWVAISIGVSPIPV--IGLRD--ESKV--FVNNTTC--VLNDPNFVLIGSFVAFFIPLTIMVITYFLTIIYVLRQTLMLLRGHTTEE--(49 a.a.)--			
D2	TRYSSKRRVTNIAIVWVLSF--TIS--CPL--L----FGLNNT--D--QNEC--IIANPAFVYSSIVSPYVP--FIVTLLVYIKIYIVLRKKRKRNTKSSR--(107 a.a.)--			
	412	VI	VII	
H1	NREKKAQKQLGFIHAAFIICWIPYFIFFNVIA--F--CESCCNQ----EVRMFTIWLGTINSTLMLIYPLCHENFKKTKKILBIRS			
H2	ICKKKAATVTLAAVMGAFIICWPFYTFVYRG--LKGDDAINE----AFEAVVLMGLTANSALMPLIYATLNRDFTYAOQLFRCRP--			
M1	VREKKAARTSAILLAFILWTPYINWLVST--F--CKDCVPE----TLWELGYLCTVNTVMPHCTASCNKAFRDHFRLLLCRW--			
$\alpha 1$	SREKKAARTLGIWVGCFLVCLWPFILVMPICGSFF--PDFRPE----TVFKIAFLWGLYLNCSIMPIIYPCSSQEFKFAFQNVLRICQ--			
5HT-1c	NREKKAQKQLGFIHAAFIICWIPYFIFFNVIA--F--CESCCNQ----EVRMFTIWLGTINSTLMLIYPLCHENFKKTKKILBIRS			
D2	QREKKAATVTLAAVMGAFIICWPFYTFVYRG--LKGDDAINE----AFEAVVLMGLTANSALMPLIYATLNRDFTYAOQLFRCRP--			

Fig. 5. Alignment of amino acid sequences of bovine histamine H₁ receptor (H1) and some representative G protein-coupled receptors. H2, canine histamine H₂ receptor (27); M1, mouse m₁-muscarinic receptor (23); $\alpha 1$, bovine $\alpha 1$ -adrenergic receptor (28); 5HT-1c, rat serotonin 1c receptor (13); and D2, rat dopamine D₂ receptor (24). Amino acid residues shown by boldfaced type in sequences are identical; residues nonhomologous with H₁ receptor sequence in the loop between transmembrane segments V–VI are summed in parentheses. Positions of putative transmembrane segments I–VII of H₁ receptor are indicated.

to be a member of the family of debrisoquine-type cytochrome P450s (38).

The receptor cDNA clone for the classical histamine receptor (3), the H₁ receptor, isolated in this study, will be useful for molecular studies of function and regulation of activities mediated through the H₁ receptor molecule and for molecular analysis of possible H₁ receptor subclasses. *In situ* and immunocytochemical studies on localization of the H₁ receptor will also be helpful in analyzing physiological functions of histamine in the central nervous system and in peripheral tissues.

We are grateful to Drs. S. Nakanishi, K. Mori, and S. Nagata for valuable advice and encouragement and to Drs. I. Imamura and H. Hayashi for help in sequence-homology analysis of receptors and hydrophathy-profile analysis of the H₁ receptor, respectively. We are also grateful to Drs. O. Hayaishi, R. Yoshida, and M. Nishizawa for valuable discussions and encouragement. This work was supported by Grants-in-Aid 63065004 and 03670102 from the Ministry of Education, Science and Culture of Japan.

- Dale, H. H. & Laidlaw, P. P. (1910) *J. Physiol. (London)* 41, 318–344.
- Ash, A. S. F. & Schild, H. O. (1966) *Br. J. Pharmacol.* 27, 427–439.
- Douglas, W. W. (1985) in *Goodman and Gilman's: The Pharmacological Basis of Therapeutics*, eds. Gilman, A. G., Goodman, L. S., Rall, T. W. & Murad, F. (MacMillan, New York), 7th Ed., pp. 605–638.
- Schwartz, J.-C., Arrang, J. M., Garbarg, M., Pollard, H. & Ruat, M. (1991) *Physiol. Rev.* 71, 1–51.
- Hill, S. J. (1990) *Pharmacol. Rev.* 42, 45–83.
- Fukui, H. (1991) in *Histaminergic Neurons: Morphology and Function*, eds. Watanabe, T. & Wada, H. (CRC, Boca Raton, FL), pp. 61–83.
- Masu, Y., Nakayama, K., Tamaki, H., Harada, Y., Kuno, M. & Nakanishi, S. (1987) *Nature (London)* 329, 836–838.
- Meyerhof, W., Schwartz, J. R., Holtt, V. & Richter, D. (1990) *J. Neuroendocrinol.* 2, 547–553.
- Sugama, K., Yamashita, M., Fukui, H., Ito, S. & Wada, H. (1991) *Jpn. J. Pharmacol.* 55, 287–290.
- Mizushima, S. & Nagata, S. (1990) *Nucleic Acids Res.* 18, 5322.
- Chomczynski, P. & Sacchi, N. (1987) *Anal. Biochem.* 152, 156–160.
- Aviv, H. & Leder, P. (1972) *Proc. Natl. Acad. Sci. USA* 69, 1408–1412.
- Julius, D., MacDermott, A. B., Axel, R. & Jassell, T. M. (1988) *Science* 241, 558–564.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463–5467.
- Cullen, B. R. (1987) *Methods Enzymol.* 152, 684–704.
- Inagaki, N., Fukui, H., Taguchi, Y., Wang, N. P., Yamatodani, A. & Wada, H. (1989) *Eur. J. Pharmacol.* 173, 43–51.
- Lehrach, H., Diamond, D., Wezney, J. M. & Boldtkew, H. (1977) *Biochemistry* 16, 4743–4751.
- Vogelstein, R. & Feinberg, A. P. (1983) *Anal. Biochem.* 132, 6–13.
- Yamashita, M., Ito, S., Sugama, K., Fukui, H., Smith, B., Nakanishi, K. & Wada, H. (1991) *Biochem. Biophys. Res. Commun.* 177, 1233–1239.
- Ruat, M., Bouthenet, M. L., Schwartz, J.-C. & Ganellin, C. R. (1990) *J. Neurochem.* 55, 378–385.
- Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* 157, 105–132.
- Lefkowitz, R. J. & Caron, M. G. (1988) *J. Biol. Chem.* 263, 4993–4996.
- Shapiro, R. A., Scherer, N. M., Habecker, B. A., Subers, E. M. & Nathanson, N. M. (1988) *J. Biol. Chem.* 263, 18397–18403.
- Bunzow, J. R., Van Tol, H. H. M., Grandy, D. K., Albert, P., Salon, J., Christie, M., Machida, C. A., Neve, K. A. & Civelli, O. (1988) *Nature (London)* 336, 783–787.
- Kozak, M. (1984) *Nucleic Acids Res.* 12, 857–872.
- Strader, C. D., Sigal, I. S., Candelore, M. R., Rands, E., Hill, W. S. & Dixon, R. A. F. (1988) *J. Biol. Chem.* 263, 10267–10271.
- Gantz, I., Schaffer, M., DelValle, J., Logsdon, C., Campbell, V., Uhler, M. & Yamada, T. (1991) *Proc. Natl. Acad. Sci. USA* 88, 429–433.
- Schwinn, D. A., Lomasney, J. W., Lorenz, W., Szklut, P. J., Fremeau, R. Y., Jr., Yang-Feng, T. L., Caron, M. G., Lefkowitz, R. J. & Cotecchia, S. (1990) *J. Biol. Chem.* 265, 8183–8189.
- Kornfeld, R. & Kornfeld, S. (1985) *Annu. Rev. Biochem.* 54, 631–664.
- Mitsuhashi, M. & Payan, D. G. (1989) *Mol. Pharmacol.* 35, 311–318.
- Volpi, M. & Berlin, R. D. (1988) *J. Cell Biol.* 107, 1533–1539.
- Dillon-Carter, O. & Chuang, D.-M. (1989) *J. Neurochem.* 52, 598–603.
- Fukui, H., Inagaki, N., Ito, S., Kubo, H., Kondoh, A., Yamatodani, A. & Wada, H. (1991) in *New Perspectives in Histamine Research*, eds. Timmerman, H. & van der Goot, H. (Birkhauser, Basel), pp. 161–180.
- Strader, C. D., Sigal, I. S. & Dixon, R. A. F. (1989) *FASEB J.* 3, 1825–1832.
- Reihardt, D. & Borchard, U. (1982) *Klin. Wochenschr.* 60, 983–990.
- Parsons, M. E. (1982) in *Pharmacology of Histamine Receptors*, eds. Ganellin, C. R. & Parsons, M. E. (Wright, Bristol, England), pp. 323–350.
- Taylor, J. E. & Richelson, E. (1982) *Eur. J. Pharmacol.* 78, 279–285.
- Fukui, H., Mizuguchi, H., Liu, Y. Q., Leurs, R., Kangawa, K., Matsuo, H. & Wada, H. (1990) *Eur. J. Pharmacol.* 183, 1727–1728.

Molecular cloning of a gene encoding the histamine H₂ receptor

(gastric acid/H₂ blockers/cimetidine/tiotidine/L cells)

IRA GANTZ*, MATTHIAS SCHÄFFER†, JOHN DELVALLE‡, CRAIG LOGSDON‡, VIRGINIA CAMPBELL†, MICHAEL UHLER§, AND TADATAKA YAMADA†‡¶

Departments of *Surgery, †Internal Medicine, ‡Physiology, and §Biological Chemistry, University of Michigan Medical Center, Ann Arbor, MI 48109

Communicated by Horace W. Davenport, October 1, 1990

ABSTRACT The H₂ subclass of histamine receptors mediates gastric acid secretion, and antagonists for this receptor have proven to be effective therapy for acid peptic disorders of the gastrointestinal tract. The physiological action of histamine has been shown to be mediated via a guanine nucleotide-binding protein linked to adenylate cyclase activation and cellular cAMP generation. We capitalized on the technique of polymerase chain reaction, using degenerate oligonucleotide primers based on the known homology between cellular receptors linked to guanine nucleotide-binding proteins to obtain a partial-length clone from canine gastric parietal cell cDNA. This clone was used to obtain a full-length receptor gene from a canine genomic library. Histamine increased in a dose-dependent manner cellular cAMP content in L cells permanently transfected with this gene, and preincubation of the cells with the H₂-selective antagonist cimetidine shifted the dose-response curve to the right. Cimetidine inhibited the binding of the radiolabeled H₂ receptor-selective ligand [*methyl*-³H]tiotidine to the transfected cells in a dose-dependent fashion, but the H₁-selective antagonist diphenhydramine did not. These data indicate that we have cloned a gene that encodes the H₂ subclass of histamine receptors.

Histamine is one of the major determinants of gastric acid secretion. On the gastric parietal cell, histamine exerts its stimulating action through an H₂ subclass of receptor coupled via a guanine nucleotide-binding protein (G protein) to activation of adenylate cyclase and production of cAMP. Antagonism of histamine's action at this receptor has been the cornerstone of an immense market for pharmacological treatment of acid-peptic disorders of the gastrointestinal tract. Through its three known receptor subclasses (H₁, H₂, and H₃), histamine has been shown to exert a broad array of other physiological actions as well, including mediation of allergic and anaphylactic responses, modulation of cardiac contractility and systemic blood pressure, and mediation of neural function in the central nervous system (1–4). Despite this wealth of pharmacological information, little is known about the structure of the histamine receptor. The present studies describing the cloning and sequencing[¶] of a gene encoding a protein with the functional characteristics of an H₂ subclass of histamine receptors provide insight into the molecular biology of histamine action.

In recent years the genes for a family of G protein-linked receptors have been cloned, and analysis of the deduced structures of their proteins has indicated that they have a motif of seven transmembrane regions. Capitalizing on the similarities of the amino acids comprising the transmembrane regions, Libert *et al.* have devised a strategy to clone other members of this family (5). By using synthetic oligonucleotides complementary to the DNA encoding the transmembrane regions of known G protein-linked receptors as primers

for the polymerase chain reaction (PCR), they were able to generate partial cDNA sequences encoding proteins having the common transmembrane motif. We utilized this strategy to clone the histamine H₂ receptor gene, using cDNA from canine gastric parietal cell mRNA as a template.

MATERIALS AND METHODS

Isolation of Parietal Cell mRNA. Cells from freshly obtained canine fundic mucosa were dispersed by sequential exposure to crude collagenase at 0.25 µg/ml and 1 mM EDTA, and a fraction enriched in parietal cells (70%) was isolated by counterflow elutriation by the method of Soll (6). RNA was extracted by the acid guanidinium isothiocyanate-phenol-chloroform method (7), and poly(A)⁺ RNA was obtained by oligo(dT)-cellulose chromatography. The poly(A)⁺ RNA served as a template for cDNA synthesis using the avian myeloblastosis virus reverse transcriptase (Seikagaku America, Rockville, MD). The cDNA thus obtained functioned as a template for the PCR with the oligonucleotide primers described below.

PCR. Oligonucleotides corresponding to the third and sixth transmembrane domains of G protein-linked receptors were duplicated from the design of Libert *et al.* (5) with the exception that our primers lacked the linker sequences. The primers were synthesized by using an Applied Biosystems 380B DNA synthesizer. The conditions for the PCR were as follows: denaturation for 1.5 min at 94°C, annealing for 2 min at 45°C, and extension for 4 min at 72°C. The reaction was carried out for 30 cycles, and then 20% of the product was added to fresh buffer and submitted to another 30 cycles. The final reaction products were extracted with phenol/chloroform, 1:1 (vol/vol), and then precipitated with ethanol. DNA polymerase I Klenow fragment was used to form blunt-ended DNA, and the products of this reaction were electrophoresed on a 2% NuSieve/1% Seaplaque gel (FMC). Of the two major bands that were produced, the one of ≈400 base pairs (bp) was cut from the gel and subcloned directly into the phage M13 sequencing vector (8). Dideoxynucleotide sequencing was then performed by the chain-termination method of Sanger (9) with Sequenase version 2 (United States Biochemical).

Genomic Cloning. The partial-length PCR-derived clone was random-primed (10) with ³²P and used as a probe to screen a canine genomic library (Clontech). Under high-stringency hybridization [0.9 M sodium chloride/0.09 M sodium citrate (6× SSC) at 65°C] and wash conditions (0.1× SSC at 55°C), a single clone exhibited a positive hybridization signal with the probe. Restriction enzyme mapping of the

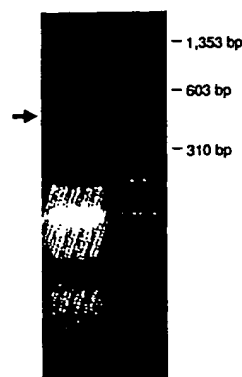
Abbreviations: G protein, guanine nucleotide-binding protein; PCR, polymerase chain reaction.

[¶]To whom reprint requests should be addressed at: 3101 Taubman Center, University of Michigan Medical Center, Ann Arbor, MI 48109–0368.

[¶]The sequence reported in this paper has been deposited in the GenBank data base (accession no. M32701).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "advertisement" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Expression Experiments. The presumed full-length coding region of the receptor was subcloned into CMVneo, a PUC13-based vector that also contains the *lacUV5*-SV40 (simian virus 40) promoter (440 bp), Tn5-neo (1400 bp) SV40 splice site and polyadenylation signal (320 bp), cytomegalovirus (CMV) promoter (700 bp), and human growth hormone polyadenylation signal (700 bp) (11). L cells were transfected by the technique of calcium phosphate coprecipitation (12). Permanently transfected L cells were selected by adding the neomycin analogue G418 to the culture medium at 600 $\mu\text{g/liter}$. The expression of the receptor gene in the selected clones was examined by RNA blot hybridization (Northern) analysis (see below) coupled with functional assays as follows. The cells were incubated in Earle's balanced salt solution with varying concentrations of histamine for 60 min at 37°C after a 60-min preincubation in medium with or without 100 μM cimetidine. Ice-cold 30% trichloroacetic acid was added to stop the reaction and precipitate the cellular protein. After centrifugation for 10 min at 1900 $\times g$, the supernatant was extracted with ether, lyophilized, and



resuspended in 50 mM Tris/2 mM EDTA, pH 7.5. The content of cAMP was measured by a competitive protein-binding assay using an Amersham kit. For binding studies, transfected L cells were plated and grown to confluence in 2.4

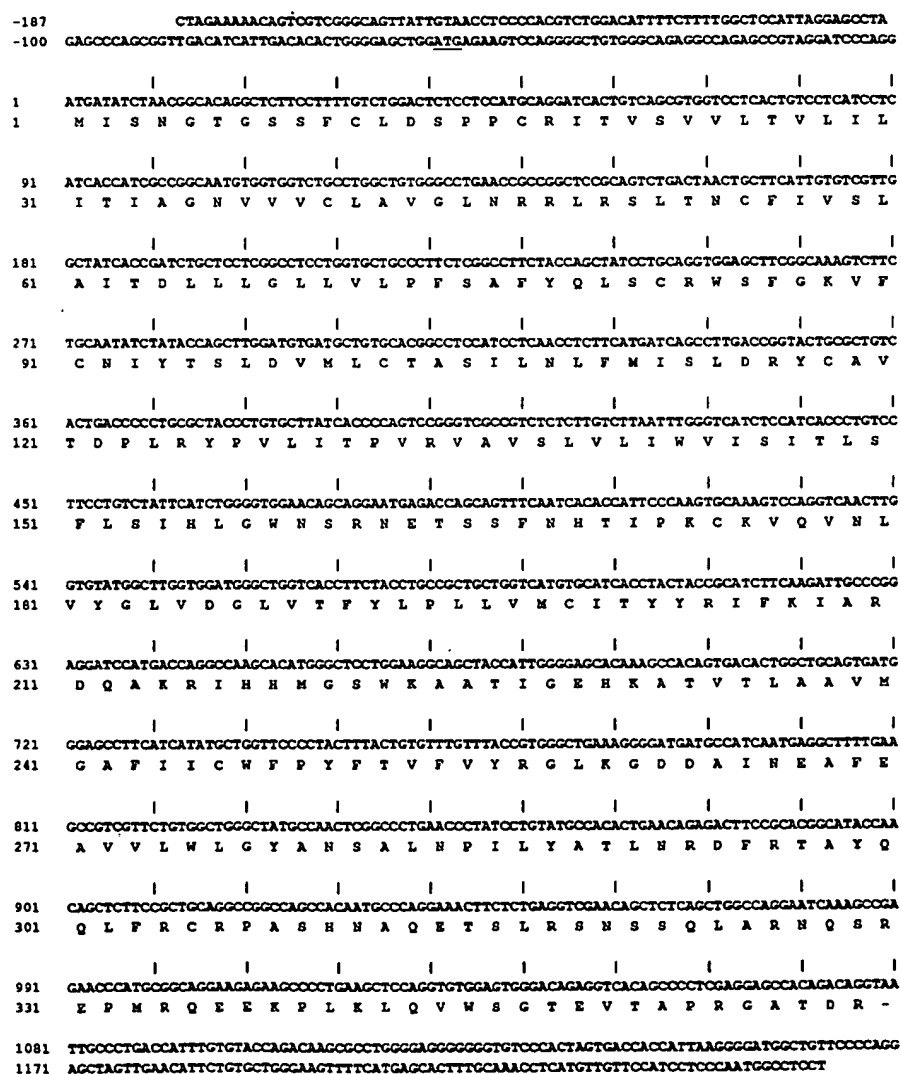


FIG. 2. The nucleotide and deduced amino acid sequence (in single-letter code) of the canine histamine H2 receptor gene.

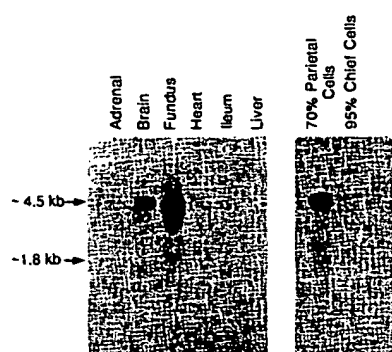


FIG. 3. Expression of the canine histamine H₂ receptor gene in various tissues. (Left) Northern blot showing the hybridization of 10 μ g of poly(A)⁺ RNA extracted from each of the designated tissues with the ³²P-labeled gene. (Right) Comparison of the expression of the receptor gene in a fraction of fundic mucosal cells consisting of roughly 70% parietal cells and a fraction consisting of the nearly 100% chief cells, the primary contaminant in the parietal cell-enriched fraction.

$\times 1.7$ cm multiwell plates. The culture medium was removed, and cells were washed twice with Earle's balanced salt solution containing 0.1% bovine serum albumin. An aliquot (36 nCi; 1 Ci = 37 GBq) of [*methyl*-³H]tiotidine (87 Ci/mmol; DuPont) was added to the culture in the presence of either cimetidine or diphenhydramine; after 1 hr of incubation, the medium was removed by aspiration. After, the cells were washed twice with phosphate buffered saline (PBS), pH 7.4, and lysed with 1% Triton X-100, the radioactivity was quantified. Maximum binding was determined by incubation of [*methyl*-³H]tiotidine with transformed L cells in the absence of antagonists. Nonspecific binding, which was subtracted from total binding to obtain specific binding, was determined as the amount of label remaining bound in the presence of 100 μ M histamine.

Northern Blots. The expression of the cloned gene was examined in various tissues by Northern blot analysis. For these studies, poly(A)⁺ RNA was extracted as described above, separated on a 1.25% formaldehyde-agarose gel, and blotted to nitrocellulose. Hybridization was performed under conditions as described (13) with the presumed coding region of the receptor gene that had been labeled with ³²P by random priming (10). The final washing of the blot was in 0.1 \times SSC at 65°C.

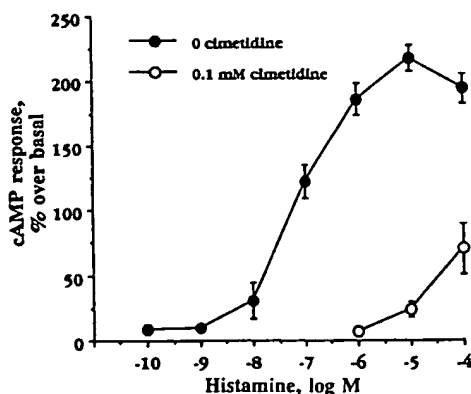


FIG. 4. Response to exogenously administered histamine of L cells transfected with a CMVneo vector containing the canine histamine H₂ receptor gene insert. The data represent means \pm SEM from four experiments. Response was shifted by addition of 0.1 mM cimetidine, an H₂ receptor-selective antagonist.

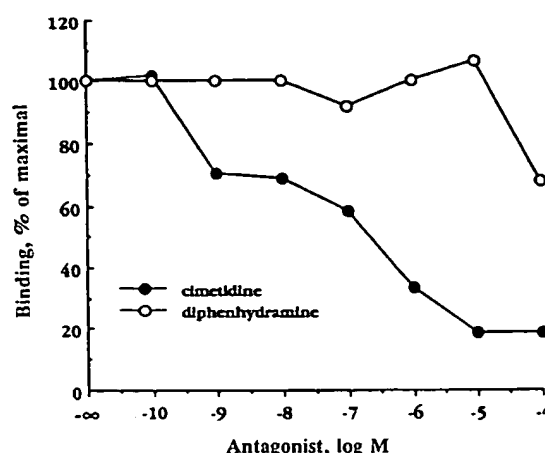


FIG. 5. Inhibition of [*methyl*-³H]tiotidine binding to transfected L cells by diphenhydramine and cimetidine. The data are from a single experiment and are virtually identical to the data obtained in two other experiments.

RESULTS

An ethidium bromide-stained gel of the products of PCR is depicted in Fig. 1. As noted above, two major bands of ≈ 400 bp and ≈ 350 bp were produced, and the former band was cut from the gel and cloned into phage M13. Of 12 clones obtained, only 1 had the nucleotide and deduced amino acid sequence expected of a G protein-linked seven-transmembrane receptor. Computer analysis of the amino acid sequence of this single clone revealed extensive homology to other known G protein-linked receptors, and Kyte-Doolittle analysis confirmed the presence of the two hydrophobic putative transmembrane domains between the third and sixth transmembrane sequences upon which the primers were based (14). Screening a canine genomic DNA library resulted in one clone with a positive hybridization signal. The nucleotide and deduced amino acid sequence of the presumed coding region of this gene is depicted in Fig. 2. Northern blot analysis showed that the gene was expressed most abundantly in the gastric fundus and, to a lesser extent, in the brain (see Fig. 3). Further analysis revealed that parietal cells were most likely to be the origin of the positive hybridization signal obtained with gastric poly(A)⁺ RNA.

The L cells transfected with the H₂ receptor construct showed dose-dependent increases in cellular cAMP content in response to histamine stimulation (Fig. 4), reaching a maximum response of $217 \pm 10\%$ over basal (mean \pm SEM; $n = 3$) after the 10 μ M histamine dose. The dose-response curve could be shifted to the right by the H₂ receptor-selective antagonist cimetidine. Serotonin, epinephrine, dopamine, and carbamoylcholine in doses as high as 100 μ M had no effect on cAMP content. Nontransfected L cells, L cells transfected with a CMVneo vector missing the receptor gene construct insert, and L cells transfected with a CMVneo vector containing as an insert a gene encoding the α catalytic subunit of the cAMP-dependent protein kinase all failed to demonstrate any response to histamine. Cimetidine displaced binding of [*methyl*-³H]tiotidine to transfected cells in a dose-dependent fashion with an ED₅₀ of $5.5 \pm 0.6 \times 10^{-7}$ M (mean \pm SEM; $n = 4$) (Fig. 5). In contrast, diphenhydramine, a relatively selective H₁ receptor antagonist, demonstrated no ability to inhibit [*methyl*-³H]tiotidine except at the highest dose.

DISCUSSION

We utilized the PCR to clone a gene encoding a protein with the functional properties of a histamine H₂ receptor. Al-

		I	
CANH2	-	MISNGTGSFCLDSPPCRTIVS-----	VV-LTVLI-----LITIAGHVVCGLVGLRRRLSL
HAMADRB2	-	HGPPGNDSDFLITNGSHVPHDVTEDDEAWVVGAILMSVIVLAIVGCF-----	GHVLVITAIKAPERLQTV
HUMADB3	-	-MAPWPHENSSLAPODPLTAPNTANTS-GLPGVWEAALAGALLAVLATV-----	NLLVIVIAIAWTPRLQTN
BOVSUBK	-	MGACVMTDINISS-GLDSNATGI---TAFSMGQWQALATAAY---LAL-VLVAVM-----	GHAIVWIILAHQRHRTV
HUMACHRM2	-	MFVNPILFPKCFKLPATWLLIRERKMNSTNSSNLSALTSPYKTFEVVFI-VLVAGSLSVTIIGHILVMVSIKVRHLQTV	
RATDOP2	-	MDPLNLSWYDDDLERQNSRPFNGSEGRADPHYNYAMLLTLLIFIVPGN-VLVCH-----	AVSREKALQTT
		II	
CANH2	-	NCPIVSLAITDLLGLLVLPFSAPYQLSCRMSPGKVPFCNIYTSLDVHC-TA---SILNLFMISLDRTYCA-VTDP-LRYPVLIT	
HAMADRB2	-	NYFITSLACADLVHGLAVVPPGASHILKMNPNFPCFPTWSIDV-LCVTA---SIETLCVIAVDRTYIAITSPFKYQS---LLT	
HUMADB3	-	NVPVTSLAAADLVHGLLVPPPAATLALTGHWPLGATGCELTWSVDV-LCVTA---SIETLCALAVDRYLA-VINT-LRYGALVT	
BOVSUBK	-	NYFIVNLALADLCAAFNAAPNFVYASHNINYPGRAPCYFQNLFPPI---TAMFVSIYSHTAIAADRTHAIVHPFPQR---LSA	
HUMACHRM2	-	NYFLFSLACADLIIGVFSMNLYTLTYVIGYWPLGPVVDLWLALDYVVS-NASVHLLIISF---DRYFC-VTKP-LTYPVKRT	
RATDOP2	-	NYLIVSLAVADLLVATLVHMPVVVLEVVGWKFPSRIHCDIFVTLDMHC-TA---SILNLCASIDRTYA-VAMPMLNTRYSS	
		III	
CANH2	-	V---RVAV---SLVLIWVISITLSPLSIHLGWNRS-NETSSFNHTIPKCKVQV-----NLVYGL-VDGLVTFTPLP---LLV	
HAMADRB2	-	NKARM-V-ILM---VMVSGLTSLPIQIMHY-RATHQKAI-----CYHKET-CCDPFTNQAY-AIASIVSFTVP---LVV	
HUMADB3	-	RCARTAV---VLVMVSAVASFAPIMSQW-RVGA---DAEAQRCHSNPR-CCAFASNMPYVL-LSSSVSFTPLP---LLV	
BOVSUBK	-	GT-R-AV-IAGINLVALALAPPQCFYSTITT-----DEGATKCVWAPEDSGGMLLLYHLIVIALIYF-LP---LVV	
HUMACHRM2	-	KMAGHM---IA---AAMVLSFILWAPAILF-WQFIVGVRTVEDGE---CYIQFFS---NAAVTFGT-AIAAFYLP---VII	
RATDOP2	-	R---RVTVMIAI---VMVLSFTISCLPLFLGNNTDQNE-----CIANPAFVVY-----SSIVSFTVPFIVTLLV	
		IV	
CANH2	-	CITYYRIFKIARDQ-----AKRIHMGSWKAATIG-----	EHKAT
HAMADRB2	-	VFVYSRVFQVEGRPHSPNL---AKRQLQKIDKSGQVEQDGRSGHGLRRSSKPKLK-----	EHKAL
HUMADB3	-	LFVTARVFPV-----ATRQLRLRGELGRFPPEESAPPSSRLAPAVGTAPPEGVPAOGRPARLLPLR---EHKAL	
BOVSUBK	-	FPVYSVIGLTLWRR-----SVPGHQAAGANLRHLQAKKKFV	
HUMACHRM2	-	T-VLYWHISV---137 aa---ARKIVKMTQPAKKKPPPSR-----	EHKVT
RATDOP2	-	IKIYIVLRKRKRKVNTRSSRAFRANLKT-----72 aa-----	FFBIQTMNGKTRTSLKTHSRRL-SQQKEKKAT
		V	
CANH2	-	TLAAVWQAFIICWFFYFTVFFYRGLEGDDAINE-AFEAV---VL-WLQYANSALNFIYATLNRDFTAYQ-QL-FRCRPAASHNA	
HAMADRB2	-	TLGIIMGTFTLCWLPFFIVNIVHVIQDNLIPKEYI---L-LNMLQYVNSAFNPLIYCRS-PDFRIAPQELL---CLRSSSK	
HUMADB3	-	TLGLIMGTFTLCWLPFFLANVLRALGGPSLVPGPAP---LALNMLQYVNSAFNPLIYCRS-PDFRSAPRRL-CRCGRRLPPE	
BOVSUBK	-	TMVLVVVTPAICWLPYHLYFILGTQ-EDYCHKFIQQVYALFVLAHSSHTYNIYICCLNHRFRSGFR---LAFRCRCPWTP	
HUMACHRM2	-	TILAILLAFIITWAPYN-VHVLINTFCAPCIPNTVNTIGY---WLCYINSTINPACALCNATPKKTFKHLH---CHYKNICA	
RATDOP2	-	MLAIVLGVPIICWLPFFITHILNHCDCNIHQSTAPSH-----GWAMSTVPSTPSTPSTPSTSSARP	
		VI	
CANH2	-	ET-----SLRSNSSLQA-RNQSREPHRQEKPLK-LQVWSGTEVTAPRGATOR	
HAMADRB2	-	YNGY-----SSNSNGKTDYNGEASGCLQG-EKESERLCEDPPGTESFVNCQGTVPSSLSDSQGRNCSTNDSPL	
HUMADB3	-	CAAAAPALFPGVPAARSSPAQRLQLDQ	
BOVSUBK	-	EDRMELTYTPSLSTRVNRCHEKEIFMSGOVAPSEAVNGQAESPAQGVSTEP	

Fig. 6. Structural comparison of the putative histamine H2 receptor with other G protein-linked receptors. The deduced amino acid sequences of the receptors (indicated by the conventional single-letter abbreviations) are aligned on the basis of homologous regions, which are shown by boldface letters. The roman numerals indicate the putative transmembrane domains. CANH2, canine H2 receptor; HAMADRB2, hamster β_2 -adrenergic receptor (15); HUMADB3, human β_3 -adrenergic receptor (16); BOVSUBK, bovine substance K receptor (17); HUMACHRM2, human M_2 -muscarinic receptor (18); RATDOP2, rat dopamine D2 receptor (19).

		A	
CANH2	-	CNTYT-SLD-VMLC-TA---SILNLFMISLDRTY	
HUMAD2	-	CGVYL-AID-VLFC-TS---SIVHLCAISLDRTY	
HUMAD1	-	CELWT-SVD-V-LCVTA---SIETLCVIAVDRTY	
HAMADRB2	-	CEFWT-SID-V-LCVTA---SIETLCVIAVDRTY	
HUMADB3	-	CELWT-SVD-V-LCVTA---SIETLCALAVDRY	
RATDOP2	-	CDIEVT-ID-VMC-TA---SILNLCASIDRTY	
HUMACHRM2	-	CDLWLA-LDYVSN-A---SVNLLIISFDRTY	
RATSUBP	-	CKFNFPIIAALF-A---SIYSMTAVAFDRY	
BOVSUBK	-	CYFQNLFPPI---TAMFVSIYSHTAIAADRTY	
MAS	-	YTVTILS---VTFEGNTGL---LLTALSVERC	
		B	
CANH2	-	NLVYGL-VGLVTFYLP---LLVMCITY	
HUMAD1	-	NRAYAI-ASSVVSFYVP---LCIMAFVY	
HAMADRB2	-	NQAYAI-ASSIVSFYVP---LVMVVFY	
HUMADB3	-	NMPYVL-LSSSVSFYLP---LLVMLFVY	
HAMADRA1	-	EPFYAL-FSSLGSEFYPLAV-ILVMYC	
RATDOP2	-	NPAFVV-YSSIVSFYVPFIVTLLVYIKIY	
BOVSUBK	-	LLVHLIVIALIYF-LP---LVMV-FVA	
RATSUBP	-	EKAYHICVTLIYF-LP---LLV-IGY	
HUMACHRM2	-	NAAVTGTATAA-FYLP---VTIMT-VL	
MAS	-	DCRAVITITALLSF-LVFTPLMNVSTIL	

Fig. 7. Structural comparisons of the third (A) and fifth (B) transmembrane domains of the canine H2 receptor (CANH2) with those of other G protein-linked receptors: HUMAD2 (26), HUMAD1 (27), and HUMADB3, human α_2 , β_1 , and β_3 -adrenergic receptors; HAMADRA1 and HAMADRB2, hamster α_1 - and β_2 -adrenergic receptors; HUMACHRM2, human M_2 -muscarinic receptor; RATDOP2, rat dopamine D2 receptor; RATSUBP (28, 29), rat substance P receptor; BOVSUBK, bovine substance K receptor; MAS, product of *mas* oncogene (30).

though the approach that we utilized to obtain this clone was nonspecific, we purposely targeted a particular tissue known to contain certain G protein-linked receptors of interest, including those for histamine and gastrin. The full-length clone obtained was initially for a receptor specific for an unknown ligand; however, comparison of the deduced amino acid sequence to that of other G protein-linked receptors with presumed seven-transmembrane motifs revealed extensive homology (Fig. 6). Like the genes encoding many of the other members of this family, our gene appeared to be devoid of introns as well (20). Several features of the amino acid sequence deduced from our gene were notable and provided clues as to its identity. The first clue was the aspartic acid residue in the third transmembrane domain. An aspartic acid in this position has been shown by mutational analysis to be important for ligand binding to the β -adrenergic receptor, which is also a member of this receptor family (Fig. 7A). It is hypothesized that the carboxyl group of the aspartic acid moiety acts as a counter anion to the cationic amino group of β -adrenergic agonists (21). Indeed, receptors for a number of cationic biogenic agonists such as dopamine and acetylcholine are also characterized by the presence of this aspartic acid residue, while receptors for other ligands such as peptide hormones are not. The second structural feature of note was the absence of the two serine residues present in the fifth transmembrane region of receptors for catecholamines and dopamine as highlighted in Fig. 7B. This information sug-

gested that our clone encoded a novel class of receptor. However, the conservative substitution of a threonine residue and an aspartic residue for the two serine residues was of particular interest in view of the data suggesting that the serines are sites of hydrogen bonding to the hydroxyl groups present in the catechol ring of adrenergic agonists (22). A third structural feature of interest (Fig. 6) was the homology of the carboxyl- and amino-terminal ends of the third cytoplasmic loop (between the fifth and sixth transmembrane regions) with comparable regions of the β_2 -adrenergic receptor, which have been shown previously to be of critical importance to its linkage to the G protein associated with adenylate cyclase activation (22, 23).

This structural information suggested the possibility that our clone encoded a receptor for a positively charged biogenic amine linked to adenylate cyclase activation. We hypothesized that the most likely such receptor on gastric parietal cells would be the H2 subtype of histamine receptor. This hypothesis was tested and proven by inserting the presumed coding region of the receptor gene into the eukaryotic expression vector CMVneo, expressing it in mouse L cells, and measuring the changes in cellular cAMP content induced by histamine. We characterized further the nature of the histamine receptor subtype encoded in our cloned gene by demonstrating the specific binding of [methyl-³H]-tiodine, a labeled H2-receptor antagonist, to L cells transformed with the receptor gene. Our data confirmed that our clone encoded the H2 subtype of histamine receptor.

An interesting feature of our cloned gene is the presence of an out-of-frame ATG codon 50 bp upstream of the presumed initiation codon of the major open reading frame (Fig. 2). A similar short open reading frame upstream of the major open reading frame has been described previously for the β -adrenergic receptor, although its significance is yet unknown (15, 24). The translation initiation sequence of the major open reading frame is more consistent with the consensus eukaryotic translation initiation sequence (25). The transcription initiation site of our receptor gene has not been determined; however, we examined two different receptor gene constructs in L cells, one containing the entire gene sequence as described in Fig. 2 and the other lacking the short upstream open reading frame. Expression of both of these constructs resulted in L cells that exhibited histamine binding and cAMP generation in response to histamine (data not shown). While we did not compare levels of expression, the upstream segment is apparently not essential for histamine receptor gene expression.

As mentioned above, a major difference in the structural features of the H2 receptor and that of catecholamine receptors is the absence of the two serine residues in the fifth transmembrane domain. However, with the knowledge that the natural ligand for the former receptor is an imidazole, it is possible to speculate on the nature of the ligand-receptor interaction. The aspartic and threonine residues that have substituted for the serine moieties have the ability to interact via hydrogen bonds with the nitrogen moieties on the imidazole ring of histamine. Future mutational analysis of this site will be required to substantiate the validity of this model. Nonetheless, through modeling and analysis it may be possible to define the nature of histamine binding and, perhaps more importantly from a therapeutic standpoint, inhibition of histamine binding.

By taking advantage of the marked homology between receptors linked to G proteins, we have been successful in cloning a gene encoding the H2 subtype of histamine receptors despite starting without even rudimentary knowledge of the biochemistry of this receptor. If there were substantial homology among the histamine receptor subtypes as there is, for example, among the catecholamine receptor subtypes, it might be possible to extend these findings on the H2 receptor

ultimately to structural information on the H1 and H3 receptors through cloning of their genes as well.

We thank Il Song for synthesizing our oligonucleotides, Jung Park for helping to provide parietal cells, Chris Dickinson for assistance in cell culture, Daryl Daugherty for general advice, and Dr. William Silen for his generous career counseling to I.G. This work was supported by National Institutes of Health Grant RO1DK34306, and RO1DK41350 and funds from the University of Michigan Gastrointestinal Peptide Research Center (National Institutes of Health Grant P30DK34933). I.G. is a recipient of a Veterans Administration Research Associate Award, and M.S. is the recipient of the Deutsche Forschungsgemeinschaft Grant Scha 453/1-1.

- Hirschowitz, B. I. (1985) in *Frontiers in Histamine Research*, eds. Ganellin, C. R. & Schwartz, J. C. (Pergamon, Oxford), p. 251.
- Pearce, F. L., Ali, H., Barret, K. E., Befus, A. D., Bienenstock, J., Brostoff, J., Ennis, M., Flint, K. C., Johnson, N. M., Leung, K. B. P. & Peachell, P. T. (1985) in *Frontiers in Histamine Research*, eds. Ganellin, C. R. & Schwartz, J. C. (Pergamon, Oxford) p. 411.
- Philippu, A. (1985) in *Frontiers in Histamine Research*, eds. Ganellin, C. R. & Schwartz, J. C. (Pergamon, Oxford) p. 335.
- Arrang, J.-M., Garbarg, M., Lancelot, J.-C., Lecomte, J.-M., Pollard, H., Robba, M., Schunack, W. & Schwartz, J. C. (1987) *Nature (London)* 327, 117-123.
- Libert, F., Parmentier, M., Lefort, A., Dinsart, C., Van Sande, J., Maehaut, C., Simons, M.-J., Dumont, J. E. & Vassart, G. (1989) *Science* 244, 569-572.
- Soll, A. H. (1978) *J. Clin. Invest.* 61, 370-380.
- Chomczynski, P. & Sacchi, N. (1987) *Anal. Biochem.* 162, 156-159.
- Crouse, G. F., Frischauf, A. & Lehrach, H. (1983) *Methods Enzymol.* 101, 78-89.
- Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
- Vogelstein, R. & Feinberg, A. P. (1983) *Anal. Biochem.* 132, 6-13.
- Brown, N. A., Steofko, R. E. & Uhler, M. D. (1990) *J. Biol. Chem.* 265, 13181-13189.
- Okayama, H. & Chen, C. (1987) *Mol. Cell. Biol.* 7, 2745-2752.
- Campbell, V. W., Del Valle, J., Hawn, M., Park, J. & Yamada, T. (1989) *Am J. Physiol.* 256, G631-G636.
- Kyte, J. & Doolittle, R. R. (1982) *J. Mol. Biol.* 157, 105-132.
- Dixon, R. A. F., Kobilka, B. K., Strader, D. J., Benovic, J. L., Dohman, H. G., Frielle, T., Bolanowski, M. A., Bennett, C. D., Rands, E., Diehl, R. E., Mumford, R. A., Slater, E. E., Sigal, I. S., Caron, M. G., Lefkowitz, R. J. & Strader, C. D. (1986) *Nature (London)* 321, 75-79.
- Emorine, L. J., Marullo, S., Briand-Sutren, M.-M., Patey, G., Tate, K., Delavie-Klutchko, C. & Strosberg, A. D. (1989) *Science* 245, 1118-1121.
- Masu, Y., Nakayama, K., Tamaki, H., Harada, Y., Kuno, M. & Nakanishi, S. (1987) *Nature (London)* 329, 836-838.
- Peralta, E. G., Ashkenazi, A., Winslow, J. W., Smith, D. H., Ramachandran, J. & Capon, D. J. (1987) *EMBO J.* 6, 3923-3929.
- Bunzow, J. R., Van Tol, H. H. M., Grandy, D. K., Albert, P., Salon, J., Christie, M., Machida, C. A., Neve, K. A. & Civelli, O. (1988) *Nature (London)* 336, 783-787.
- Kobilka, B. K., MacGregor, C., Daniel, K., Kobilka, T. S., Caron, M. G. & Lefkowitz, R. J. (1987) *J. Biol. Chem.* 262, 15796-15802.
- Strader, C. D., Sigal, I. S., Candelore, M. R., Rands, E., Hill, W. S. & Dixon, R. A. F. (1988) *J. Biol. Chem.* 263, 10267-10271.
- Strader, C. D., Sigal, I. S. & Dixon, R. A. F. (1989) *FASEB J.* 3, 1825-1832.
- Lefkowitz, R. J. & Caron, M. G. (1988) *J. Biol. Chem.* 262, 4993-4996.
- Allen, J. M., Abrass, I. B. & Palmiter, R. D. (1989) *Mol. Pharmacol.* 36, 248-255.
- Kozak, M. (1983) *Nucleic Acids Res.* 12, 857-872.
- Kobilka, B. K., Matsui, H., Kobilka, T. S., Yang-Feng, T. L., Francke, U., Caron, M. G., Lefkowitz, R. J. & Regan, J. W. (1987) *Science* 238, 650-656.
- Frielle, T., Collins, S., Daniel, K. W., Caron, M. G., Lefkowitz, R. J. & Kobilka, B. K. (1987) *Proc. Natl. Acad. Sci. USA* 84, 7920-7924.
- Yokota, Y., Sasai, Y., Tanaka, K., Fujiwara, T., Tsuchida, K., Shigemoto, R., Kakizuka, A., Ohkubo, H. & Nakanishi, S. (1989) *J. Biol. Chem.* 264, 17649-17652.
- Hershey, A. D. & Krause, J. E. (1990) *Science* 247, 958-962.
- Jackson, T. R., Blair, L. A. C., Marshall, J., Goedert, M. & Hanley, M. R. (1988) *Nature (London)* 335, 437-440.

ACCELERATED COMMUNICATION

Cloning and Functional Expression of the Human Histamine H₃ Receptor

TIMOTHY W. LOVENBERG, BARBARA L. ROLAND, SANDY J. WILSON, XIAOXIA JIANG, JAYASHREE PYATI, ARNE HUVAR, MICHAEL R. JACKSON, and MARK G. ERLANDER

R.W. Johnson Pharmaceutical Research Institute, San Diego, California

Received February 12, 1999; accepted April 2, 1999

This paper is available online at <http://www.molpharm.org>

ABSTRACT

Histamine regulates neurotransmitter release in the central and peripheral nervous systems through H₃ presynaptic receptors. The existence of the histamine H₃ receptor was demonstrated pharmacologically 15 years ago, yet despite intensive efforts, its molecular identity has remained elusive. As part of a directed effort to discover novel G protein-coupled receptors through homology searching of expressed sequence tag databases, we identified a partial clone (GPCR97) that had significant homology to biogenic amine receptors. The GPCR97 clone was used to probe a human thalamus library, which resulted in the isolation of a full-length clone encoding a putative G protein-coupled receptor. Homology analysis showed the highest similarity to M2 muscarinic acetylcholine receptors and overall low homology to all other biogenic amine receptors. Transfection of

GPCR97 into a variety of cell lines conferred an ability to inhibit forskolin-stimulated cAMP formation in response to histamine, but not to acetylcholine or any other biogenic amine. Subsequent analysis revealed a pharmacological profile practically indistinguishable from that for the histamine H₃ receptor. In situ hybridization in rat brain revealed high levels of mRNA in all neuronal systems (such as the cerebral cortex, the thalamus, and the caudate nucleus) previously associated with H₃ receptor function. Its widespread and abundant neuronal expression in the brain highlights the significance of histamine as a general neurotransmitter modulator. The availability of the human H₃ receptor cDNA should greatly aid in the development of chemical and biological reagents, allowing a greater appreciation of the role of histamine in brain function.

Since its first pharmacological description as an endogenous substance in 1910 (Barger and Dale, 1910), histamine has proven to exert tremendous influence over a variety of physiological processes. Most notable are its roles in the inflammatory "triple response" and in gastric acid secretion, which are mediated by H₁ (Ash and Schild, 1966) and H₂ (Black et al., 1972) receptors, respectively. In the early 1970s emerged an understanding that histamine is a neurotransmitter in the central nervous system (Schwartz et al., 1970; Baudry et al., 1975). In 1983, a third subtype of histamine receptor, H₃, was identified as a presynaptic autoreceptor on histamine neurons in the brain controlling the stimulated release of histamine (Arrang et al., 1983). Subsequently, the H₃ receptor has been shown to be a presynaptic heteroreceptor in nonhistamine-containing neurons in both the central and peripheral nervous systems (for review, see Hill et al., 1997). Through the molecular cloning of H₁ and H₂, these receptors were proven to belong to the superfamily of G protein-coupled receptors (GPCRs; Gantz et al., 1991; Ya-

mashita et al., 1991). For the past 10 years, the histamine H₃ receptor has been the target of numerous cloning and purification attempts, yet its molecular identity has remained an enigma.

We have initiated an effort to identify and clone orphan GPCRs as a means to identify novel drug targets and as a way to discover novel neurotransmitters and peptides. This is an approach used by many investigators, and it has led to the successful identification of ligands such as nociceptin (Reinscheid et al., 1995), prolactin-releasing factor (Hinuma et al., 1998), the orexins (Sakurai et al., 1998), and, more recently, apelin (Tatemoto et al., 1998). There are at least 70 orphan GPCRs in the public domain. We have identified, through searching public and private databases, at least 30 additional putative members of this family via expressed sequence tags (ESTs). One of these orphan receptors, our designation GPCR97, was expressed abundantly in the central nervous system, and its 5'-most sequence shares significant homology with the putative transmembrane domain

VII of several members of the biogenic amine family of receptors. Therefore, we investigated the possibility that the GPCR97 cDNA encodes a novel neurotransmitter receptor.

Experimental Procedures

Materials. Human mRNA and all Northern blots were purchased from Clontech (Palo Alto, CA). cDNA synthesis kits were purchased from Gibco Life Technologies (Gaithersburg, MD). Gelzyme was obtained from Invitrogen (San Diego, CA), and pCIneo vector was obtained from Promega (Madison, WI). All cell lines were obtained from American Type Culture Collection (Manassas, VA). Cyclic AMP (cAMP) Flashplates were obtained from DuPont/New England Nuclear (Boston, MA). Fluo-3 was purchased from TEF Laboratories (Austin, TX) G418 was purchased from Calbiochem (San Diego, CA). All histamine ligands were purchased from Research Biochemicals, Inc. (Natick, MA). All other reagents were purchased from Sigma Chemical Co. (St. Louis, MO).

Cloning of GPCR97 cDNA. A human thalamus cDNA library was constructed from poly(A)⁺-selected RNA as described by the manufacturer (Gibco Life Technologies). Double-stranded DNA was digested with *NotI* and then run on a 0.8% low-melting agarose gel, and cDNA in the range of 2.5 to 5 kilobases (kb) was excised, purified with Gelzyme, and subsequently was subcloned into pSport vector. The size-selected human thalamus cDNA library was screened with a radiolabeled fragment of the GPCR97 EST clone. A full-length GPCR97 was obtained and, subsequently, cloned into the mammalian expression vector pCIneo (Promega) and transfected into human embryonic kidney 293 cells, rat C6 glioma cells, and human SK-N-MC neuroblastoma cells.

Transfection of Cells with GPCR97 cDNA. Cells were grown to about 70% to 80% confluence and then removed from the plate with trypsin and pelleted in a clinical centrifuge. The pellet was then resuspended in 400 μ l of complete media and transferred to an electroporation cuvette with a 0.4-cm gap between the electrodes (no. 165–2088; Bio-Rad Laboratories, Hercules, CA). One microgram of supercoiled DNA was added to the cells and mixed. The voltage for the electroporation was set at 0.25 kV and the capacitance was set at 960 μ F. After electroporation, the cells were diluted into 10 ml of complete media and were plated onto four 10-cm dishes at the following ratios: 1:20, 1:10, 1:5, and the remaining cells. The cells were allowed to recover for 24 h before the addition of G-418. Colonies that survived selection were grown and tested. Several different cell lines were used for transfection, which served two purposes. First, because single-cell cloning can often uncover endogenously expressed receptors (unpublished observations), it is imperative to see the desired function in multiple transfections in different cell lines. Second, each cell line has a unique characteristic that can be used to enhance different aspects of the study. For example, C6 cells grow very fast and are easy to culture and, thus, are good for generating lots of membranes for binding. SK-N-MC cells give robust cAMP accumulation and give efficient coupling for inhibition of adenylate cyclase. L cells consistently transfect well and have few endogenous receptors, and, thus, are good for reliable initial characterization of recombinant receptors. It should be noted that inhibition of adenylate cyclase and [³H]*R*- α -methylhistamine binding were observed in all of the GPCR97-transfected cells. Only the best responding cell lines were used for further study.

cAMP Accumulation. Transfected cells were plated on 96-well plates. Overnight cultures were then incubated with Dulbecco's modified Eagle's medium-F12 media containing isobutylmethylxanthine (2 mM) for 20 min, treated with agonists, antagonists, or both for 5 min, and then treated with forskolin (10 μ M) for 20 min. The reaction was stopped with 1/5 volume 0.5 N HCl. Cell media were then tested for cAMP concentration by radioimmunoassay with cAMP Flashplates.

Calcium Mobilization. Transfected cells were plated on black 96-well plates with clear bottoms. Overnight cultures were then

incubated with Dulbecco's modified Eagle's medium-F12 media containing the fluorescent calcium indicator fluo-3 (4 μ M) and probenid (2 mM) for 60 min. Ligand-induced fluorescence was then measured on a Fluorometric Imaging Plate Reader (FLIPR; Molecular Devices, Sunnyvale, CA).

***R*- α -Methyl[³H]histamine Binding.** Cell pellets from GPCR97-expressing C6 cells were homogenized in 20 mM Tris-HCl/0.5 mM EDTA. Supernatants from a 800g spin were collected and recentrifuged at 30,000g for 30 min. Pellets were rehomogenized in 50 mM Tris/5 mM EDTA (pH 7.4). Membranes were incubated with 0.4 nM *R*- α -methyl[³H]histamine plus/minus test compounds for 45 min at 25°C and harvested by rapid filtration over GF/C glass fiber filters (pretreated with 0.3% polyethylenimine), followed by four washes with ice-cold buffer. Nonspecific binding was defined with 10 μ M histamine. *pK_i* values were calculated based on a *K_d* of 150 pM and a ligand concentration of 400 pM (Cheng and Prusoff, 1973).

In Situ Hybridization. Three adult male Sprague-Dawley rats were perfused with 4% paraformaldehyde in 0.1 M borate buffer fixative, and their brain tissues were postfixed overnight in fixative with 10% sucrose and frozen in dry ice. Five 1-in-5 series of 30- μ m-thick coronal sections of the whole brain were cut on a sliding microtome and mounted onto glass slides. In situ hybridization was performed with ³⁵S-riboprobes on this tissue by an adapted protocol (Simmons et al., 1989). Then the tissue samples were put on X-ray film for 1 day, after which they were dipped in NBT2 nuclear emulsion (Eastman Kodak Co., Rochester, NY), and kept desiccated in the dark at 4°C for 6 days. Slides were developed, were Nissl stained, and were studied under the microscope to identify structures labeled with the GPCR97 cRNA probe.

RNA Probes. The cRNA probe was constructed from a partial rat GPCR97 cDNA clone originally identified by polymerase chain reaction (PCR) amplification from rat brain cDNA with primers designed against the human receptor (5' primer, 5'-AGTCCGATCCAGCTACGACCGCTTCCTGTC-3'; 3' primer, 5'-AGTCAAGCTTGGAGCCCTCTTGAGTGAGC-3'). The resulting 607-base pair (bp) fragment was ligated into pBluescript (Stratagene, La Jolla, CA). ³⁵S-UTP-labeled antisense and sense probes for rat GPCR97 were synthesized after linearization with *Bam*HI or *Hind*III with T7 or T3 RNA polymerase, respectively. The labeled sense strands served as controls and did not show any specific labeling of cellular localization (data not shown). Specific activities of ³⁵S-UTP probes were approximately 2 to 3 $\times 10^6$ counts per minute/ μ g. All restriction enzymes and phage RNA polymerases were obtained from Boehringer Mannheim (Indianapolis, IN).

Northern Blot Analysis. Northern blots obtained from Clontech (Palo Alto, CA) were hybridized with α -³²P-dCTP-labeled (Amersham Pharmacia Biotech, Piscataway, NJ) human GPCR97 cDNA as described by the manufacturer (Expresshyb, Clontech). Two million counts per milliliter was used in a total volume of 10 ml of hybridization buffer and incubated at 68°C for 2 h. The blot was then washed two times at RT in 2 \times standard saline citrate and 0.05% SDS for 30 min each. It was further washed two more times for 30 min each at 60°C and exposed overnight to film.

Results

Cloning and Sequence Analysis of GPCR97 cDNA. GPCR97 was initially identified as an EST in a basic local alignment search tool (Altschul et al., 1990) search of the Life Seq database (Incyte Pharmaceuticals, Palo Alto, CA) with the α_2 -adrenergic receptor sequence as a query. The 5' end of the GPCR97 EST had approximately 35% homology to the seventh transmembrane domain of the α_2 -adrenergic receptor. Semiquantitative PCR of GPCR97 with cDNA templates from a variety of human tissues showed expression predominantly in the central nervous system, with the greatest intensity in the thalamus. Therefore, we constructed a size-

region with low homology (20–27%) to the biogenic amine subfamily of GPCRs. Most notable was an aspartic acid residue in the putative transmembrane domain III, the putative binding site for the primary amine, which is a clear hallmark of the biogenic amine receptor subfamily (Fig. 1). This conserved aspartic acid residue is shown in the alignment of the

TM1

H1 MSLPN-----SSCLEDKMCCEGNKTTMASPQLMPLVVVLSTICLVTVGLNLLVLYAVR
H2 MAPNG-----TASSFCL-DSTAC--K-----ITITVVLAVLILITVAGNVVVCLAVG
GPCR97 MERAPPDGPLNASGALAGDAAAAGGARGFSAAWTAVLAALMALLIVATVLGNALVMLAFV
* * * * *

TM2

H1 SERKLHTVGNLIVSLSVADLIVGAVVMPMNILYLLMSKWSLGRPLCLFWLSMDYVASTA
H2 LNRRLRNLTNCFIVSLAITDLLGLLVLPFSAIYQLSCKWSFGKVFCNIYTSLDVMLCTA
GPCR97 ADSSLRTQNFFLLNLAISDFLVGAFCIPLYVPYVLTGRWTFGRGLCKLWLVVDYLLCTS
* * * * *

TM3

H1 SIFS VFILCIDRYRSVQQPLRYLK YRTKTR-ASATILGAWFLSFLWVIP--ILGWNHFMQ
H2 SILNLFMISLDRYCAVMDPLRYPVLVTPVR-VAISLVLIWVISITLSFLSIHLGWNSRNE
GPCR97 SAFNIVLISYDRFLSVTRAVSYRAQQGDTRRAVRKMLLVVWVLAFLLYGP-AILSWEYLSG
* ** * * *

TM4

H1 SIFS VFILCIDRYRSVQQPLRYLK YRTKTR-ASATILGAWFLSFLWVIP--ILGWNHFMQ
H2 SILNLFMISLDRYCAVMDPLRYPVLVTPVR-VAISLVLIWVISITLSFLSIHLGWNSRNE
GPCR97 SAFNIVLISYDRFLSVTRAVSYRAQQGDTRRAVRKMLLVVWVLAFLLYGP-AILSWEYLSG
* ** * * *

TM5

H1 QTSVRRED-KCETDFYDVTWFKVMTAIINFYLP TLLMLWFYAKIYKAVRQHCQHRELINR
H2 TSKGNHTTSKCKVQVNEV--YGLVDGLVTFYLP LLLIMCITYYRIFKVARDQAKR---INH
GPCR97 GSSIPEGH--CYAEFFYNWYFLITASTLEFFT PFLSVTFFNLSIYLNII--Q-RRTRLRLD
* * * *

H1 SLPSFSEIKLRPENPKGDAKKPGKESPEWEVLKRKPKDAGGGSVLKSPSQTPKEMKSPVVF
H2 ISSWKAATIREH-----
GPCR97 GAREAAAGPEPPPEAQSPPPPPGCWGCWQKGHGEAMPLHRYGVGEAAVGAEEAGEATLGGG
*

H1 SQEDDREVDKLYCFPLDIVHMQAAAEGSSRDYVAVNRSHGQLKTDEQGLNTHGASEISED
H2 -----
GPCR97 GGGGSVASPTSSSGSSSRGTERPRSLKRGSKPSASSASLEKRMKMVSQSFTQRFRLSRDR

H1 QMLGDSQSFSRSDSDTTTETAPGKGKLRSGSNTGLDYIKFTWKRLRSHSRQYVSGLHMNR
H2 -----
GPCR97 -----

TM6

H1 ERKAAKQLGFIMAAFILCWIPIYFIFFMVIAFCKNCCNEHL-----
H2 --KATVTLAAVMGAFIICWFPYFTAFVYRGLRGDDAINEVLEAIVNASQLSRTQSREPRQ
GPCR97 --KVAKSLAVIVSIFGLCWAPYTLMIIRAACHGHCVDPDYW-----
* * * * *

TM7

H1 -----HMFTIWLGYINSTLNPLIYPLCNENFKKTFKRILHIRS
H2 QEEKPLKLQVWSGTEVTAPQGATDRLWLGYANSALNPILYAALNRDFRTGYQQLFCCRL
GPCR97 -----YETSFWLLWANSAVNPVLYPLCHHSFRRRAFTKLLCPQK
* * * *

H1 -----
H2 ANRNSHKTSLRS--
GPCR97 LKTIOPHSSLEHCWK

Fig. 1. Amino acid sequence of human GPCR97 receptor compared with the human histamine H₁ and H₂ receptors. Putative transmembrane domains are stated above the sequence and indicated by a solid line. Residues that are identical among all three receptors are indicated by an * below the sequence. DNA and protein sequences have been deposited with GenBank (accession no. AF140538)

predicted amino acid sequence of GPCR97 with the human histamine H_1 and H_2 receptors. Overall homology between GPCR97 and the H_1 and H_2 receptors is 22% and 21.4%, respectively.

GPCR97-Expressing Cells Inhibit Adenylate Cyclase in Response to Histamine. Given the homology of GPCR97 to the biogenic amine family, we first tested its ability to respond to several of the amine neurotransmitters, measuring either the stimulation of calcium mobilization or the increase or decrease of cAMP accumulation in mouse L cells. The biogenic amine ligands tested (acetylcholine, dopamine, imidazole, epinephrine, tryptamine, serotonin, and histamine) were negative for an increase in both calcium mobilization or in cAMP accumulation (not shown). However, after forskolin stimulation of basal cAMP accumulation, there was a selective and marked inhibition of adenylate cyclase in response to histamine in the transfected cell line but not in the nontransfected cell line (Fig. 2). This effect was mimicked by the high-affinity H_3 agonist R - α -methylhistamine, which has an EC_{50} of 1 nM (Fig. 3). In addition, the effect of R - α -methylhistamine could be blocked by the known selective H_3 antagonists thioperamide and clobenpropit (Fig. 3) but not by the H_1 antagonist diphenhydramine (Fig. 3) or the H_2 antagonist ranitidine (not shown).

GPCR97-Expressing Cells Bind the High-Affinity Histamine H_3 Ligand R - α -Methyl[3H]histamine. To confirm the H_3 pharmacology, we examined whether the GPCR97-transfected cells could bind the H_3 ligand R - α -methyl[3H]histamine. For these studies, we transfected a different cell line (C6 glioma cells) because of its ability to grow fast. C6 cells transfected with GPCR97 were able to bind [3H] R - α -methylhistamine with high affinity (Fig. 4, inset), whereas untransfected cells had no demonstrable binding (not shown). In addition, the known H_3 agonists (histamine, imetit, and N -methylhistamine) and antagonists (thioperamide and clobenpropit) could all compete for bind-

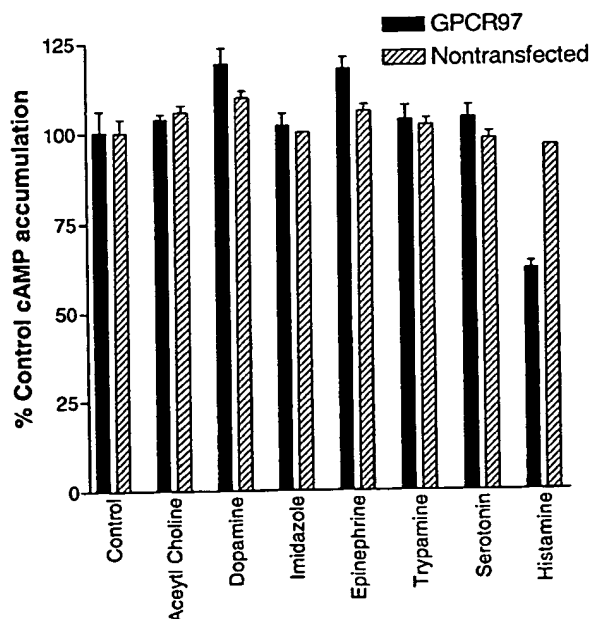


Fig. 2. Inhibition of cAMP accumulation in response to the various amine transmitters. Cells were treated with 10 μ M forskolin 5 min after the addition of compounds (1 μ M) and incubated for an additional 20 min. All values were determined in duplicate. Error bars represent S.E.M.

ing (Fig. 4) with a rank order of potency consistent with that described for the histamine H_3 receptor (Table 1).

GPCR97 is Expressed Abundantly in the Central Nervous System. Because the pharmacological profile of GPCR97 was consistent with the histamine H_3 receptor, we investigated the mRNA distribution and compared it to the known distribution of H_3 binding sites. Northern blots of human mRNA showed expression only in the brain, most notably in the thalamus and the caudate nucleus (Fig. 5). Little expression was observed in any peripheral tissue examined (heart, placenta, lung, liver, skeletal muscle, kidney, pancreas, spleen, thymus, prostate, testis, ovaries, small intestine, colon, stomach, thyroid, lymph node, trachea, and bone marrow; data not shown). To obtain a rat homolog of the GPCR97 cDNA, we used oligonucleotide primers designed from the human sequence to amplify a cDNA fragment from RNA extracted from rat brain. This rat cDNA probe (which has 85% nucleotide identity to human GPCR97) was subsequently used to examine the tissue distribution of GPCR97-encoded mRNA by in situ hybridization in rat brain sections. GPCR97 mRNA is abundantly expressed in rat brain and is most notably observed throughout the thalamus, the ventromedial hypothalamus, and the caudate nucleus (Fig. 6, A and B). Strong expression was also seen in layers II, V, and VIb of the cerebral cortex, in the pyramidal layers (CA1 and CA2) of the hippocampus, and in olfactory tubercle (Fig. 6, A and B). Because the H_3 receptor functions as an inhibitory presynaptic receptor, it is expected that the mRNA localization may not exactly match the functional receptor localization, depending on the axonal length of the neuron expressing it. For example, noradrenergic cells in the locus ceruleus project to all areas of the cerebral cortex where histamine, via H_3 receptors, is known to regulate noradrenaline release (Schlicker et al., 1989; Smits and Mulder, 1991). Therefore, it was predicted and confirmed that the mRNA for GPCR97 was expressed in the locus ceruleus (Fig. 6, C and E). In addition, because the H_3 receptor has also been functionally demonstrated on the histamine terminals in the cerebral cortex (Arrang et al., 1983), its mRNA must also be located in the histaminergic cell bodies in the tuberomammillary nuclei. This was also confirmed for GPCR97 (Fig. 6D).

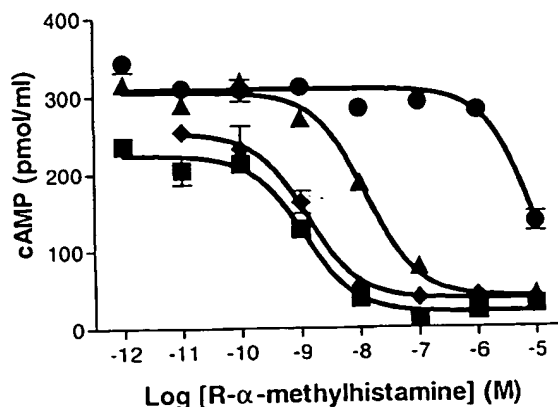


Fig. 3. Inhibition of cAMP accumulation in response to the agonist R - α -methylhistamine. Cells were treated with 10 μ M forskolin 5 min after the addition of R - α -methylhistamine and incubated for an additional 20 min. Where indicated, antagonists (1 μ M) were incubated 5 min before the addition of the agonist alone (■), with diphenhydramine (◆), with thioperamide (▲), or with clobenpropit (●). All values are determined in triplicate. Error bars represent S.E.M.

There are numerous reports of presynaptic H₃ receptors in the autonomic nervous system controlling neurotransmitter release in the heart, the lung, and the gastrointestinal tract (Arrang et al., 1988; Molderings et al., 1992; Bertaccini and Coruzzi, 1995; Imamura et al., 1995; Stark et al., 1996a). GPCR97 mRNA was detected by PCR amplification in RNA extracted from human small intestine, testis, and prostate

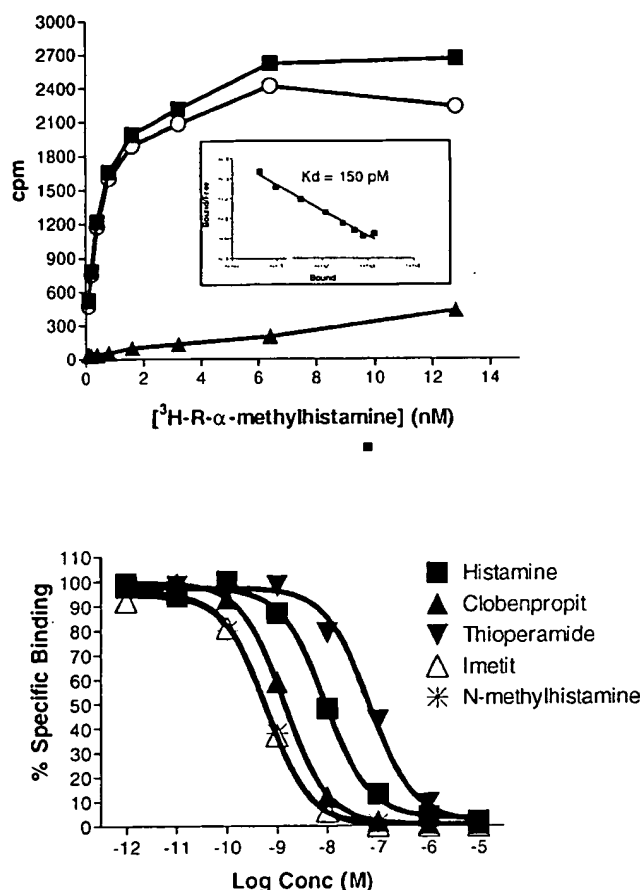


Fig. 4. Top, saturation isotherm and Scatchard transformation (inset) of *R*-α-methyl[³H]histamine to GPCR97-transfected C6 cells. Total binding (■), nonspecific binding (▲), and specific binding (○) are shown. Bottom, competition binding of [³H]*R*-α-methylhistamine (0.4 nM) in the presence of various concentrations of H₃ agonists and antagonists. *K*_D was calculated as $-1/\text{slope}$ from the linear Scatchard transformation. *p**K*₅₀ values were determined by a single site curve fitting program (Prism; GraphPad Software, San Diego, CA) and converted to *pK*₁ values according to Cheng and Prusoff (1973).

TABLE 1
*pK*₁ values of known histamine agonists and antagonists

Compound	<i>pK</i> ₁
<i>N</i> -methylhistamine	-9.8
Imetit	-9.7
Immepip	-9.7
Clobenpropit	-9.3
Histamine	-8.5
Thioperamide	-7.7
Ranitidine	>-5
Diphenhydramine	>-5
Clozapine	>-5
Cirazoline	>-5
Mepyramine	>-5
Imidazole	>-5

Values were determined by competition binding with *R*-α-methyl[³H]histamine to GPCR97-expressing cell membranes.

tissues, but was not detected in these tissues by Northern blot analysis (not shown). If GPCR97 was only expressed in the neuronal plexus, its overall low abundance in a whole tissue preparation could account for this discrepancy. We are currently investigating via in situ hybridization whether the GPCR97 receptor mRNA is produced in the ganglia of the autonomic and enteric nervous systems. An alternative explanation for the absence of clear peripheral expression could be the existence of additional subtypes of the H₃ receptor, which previously has been suggested based on pharmacological evidence (West et al., 1990; Raible et al., 1994; Leurs et al., 1996; Schlicker et al., 1996).

Discussion

The present data describes the cloning and characterization of a novel GPCR, GPCR97, with a pharmacology and a tissue distribution that is consistent with the histamine H₃ receptor subtype. We found that cells transfected with GPCR97 were able to inhibit adenylate cyclase in response to histamine. Because the two known cloned histamine receptors, H₁ and H₂, activate phosphoinositide hydrolysis and stimulation of adenylate cyclase, respectively, the inhibition of adenylate cyclase that we observed is a new finding for a cloned histamine receptor. It should be noted that previous experiments with pertussis toxin- and histamine-stimulated ³⁵S-GTPγS binding have suggested that the H₃ receptor might be G_i-linked (Clark et al., 1993; Laitinen and Jokinen, 1998). Because the putative H₃ histamine receptor has been pharmacologically defined (Arrang et al., 1987; Leurs et al., 1998), we were able to test known selective agonists and antagonists. The selective H₃ agonist *R*-α-methylhistamine was able to potently and dose-dependently inhibit forskolin-stimulated adenylate cyclase, an effect that was mimicked by two additional H₃ agonists, imetit and *N*-α-methylhistamine (data not shown). In addition, the effect of *R*-α-methylhistamine was blocked by the selective H₃ antagonists thioperam-

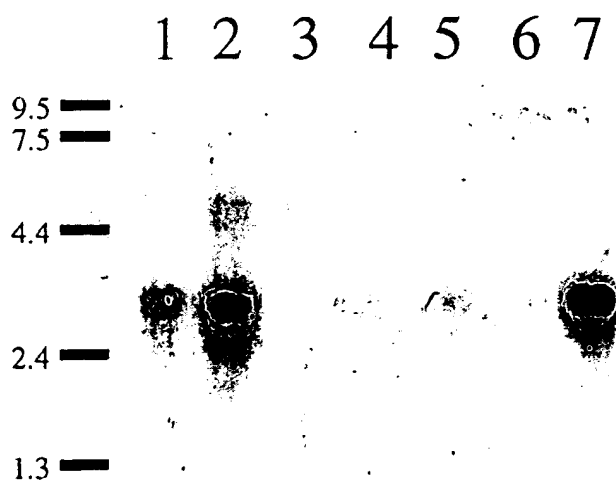


Fig. 5. Northern blot analysis of human brain mRNA samples (5 ug of poly(A)⁺ RNA/lane). Lane 1, amygdala. Lane 2, caudate. Lane 3, corpus callosum; Lane 4, hippocampus. Lane 5, whole brain. Lane 6, substantia nigra. Lane 7, thalamus. The probe was the full-length GPCR97 coding sequence. Exposure time to film was 3 days (-80°C).

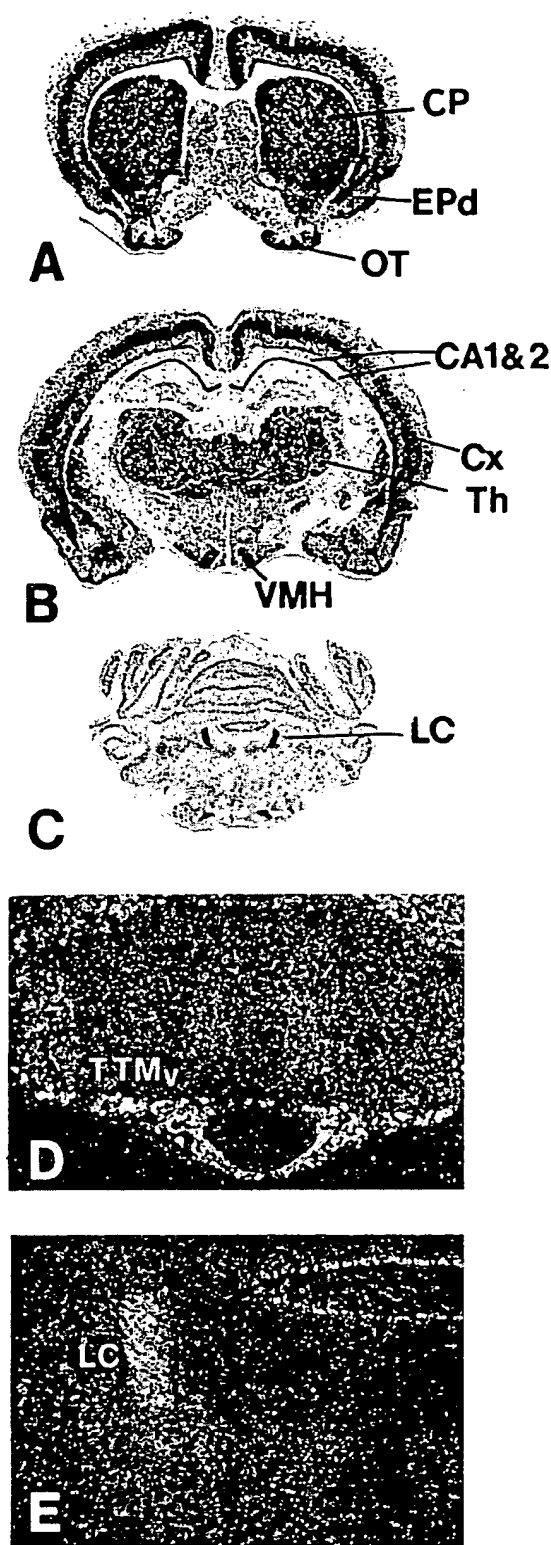


Fig. 6. Distribution of GPCR97 mRNA in rat brain. Representative film autoradiographs of coronal sections arranged rostral to caudal (A-C) and darkfield photomicrographs of coronal brain sections showing GPCR97 mRNA in the ventral portion of the tuberomammillary nucleus (D), and in the locus ceruleus (E). Magnification, D = 100 \times and E = 40 \times . Abbreviations: CA1, CA2, pyramidal layers of the hippocampus; CP, caudoputamen; Cx, cortex; EPd, endopiriform nucleus, dorsal part; LC, locus ceruleus; OT, olfactory tubercle; Th, thalamus; TTMv, tuberomammillary nucleus, ventral portion; VMH, ventromedial hypothalamus.

ide and clobenpropit but not by the H_1 or H_2 antagonists diphenhydramine or ranitidine. GPCR97-transfected cells also bound the high-affinity H_3 agonist R - α -methyl[3H]histamine. All of the tested H_3 agonists and antagonist could compete for specific R - α -methyl[3H]histamine binding with similar potencies to those reported for these compounds to brain membranes (Hill et al., 1997). It has been suggested that clozapine may impart some of its antipsychotic effects in humans through H_3 receptor antagonism (Kathmann et al., 1994; Rodrigues et al., 1995; Stark et al., 1996b). We found that clozapine did not significantly compete for binding to the recombinant human receptor (Table 1). These differences in pharmacology may be because of species differences or possible H_3 heterogeneity (West et al., 1990).

One of the most striking features of this receptor is the abundant expression in the central nervous system, particularly in the caudate, the thalamus, and the cortex. Thus, it is surprising that this receptor cDNA has eluded so many cloning attempts over the years. To explain the previous unsuccessful attempts to clone the H_3 receptor, we compared the sequence of GPCR97 to that of the H_1 and H_2 receptors (Fig. 1). The low overall homology among these three receptors suggests, in retrospect, that low-stringency hybridization approaches or degenerate PCR would not have been fruitful. In addition, we searched the public EST databases with the entire H_3 receptor mRNA sequence. We found that the H_3 receptor exists in the public domain in several clones derived from human brain libraries. However, all of these clones primarily contain only a 3'-untranslated sequence, suggesting that there may be some secondary structure present that prevents a full-length H_3 encoding mRNA from being efficiently copied by reverse transcription. Our success in screening the human thalamus may be due to its abundance in that specific brain region, coupled with the fact that we size-selected for mRNAs greater than 2.5 kb.

There are many questions that remain to be answered about the histamine H_3 receptor that we can now begin to answer with the cDNA. For example, are there additional H_3 receptor subtypes? What additional neurotransmitter systems are regulated by histamine H_3 receptors? Are H_3 receptors expressed on nonneuronal cells in the periphery? We are currently seeking to answer some of these questions. In addition, we are inactivating the H_3 receptor gene in mice (i.e., knockout mice) to identify its role in central nervous system function and memory control and as a means to look for additional phenotypes, which may lead to a better understanding of the physiological role of H_3 receptors in normal and pathological states.

Acknowledgments

We thank Jose Galindo for his great help in assembling the sequence information and K.C. Joy for performing reverse transcription-PCR experiments. We also thank Drs. Lars Karlsson, Nigel Shankley, and Josee Leysen for providing insightful discussion.

References

- Altschul SF, Gish W, Miller W, Myers EW and Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403-410.
- Arrang JM, Devaux B, Chodkiewicz JP and Schwartz JC (1988) H_3 -receptors control histamine release in human brain. *J Neurochem* 51:105-108.
- Arrang JM, Garbarg M, Lancelot JC, Lecomte JM, Pollard H, Robba M, Schunack W and Schwartz JC (1987) Highly potent and selective ligands for histamine H_3 -receptors. *Nature (London)* 327:117-123.
- Arrang JM, Garbarg M and Schwartz JC (1983) Autoinhibition of brain histamine

- release mediated by a novel class (H₃) of histamine receptor. *Nature (London)* 302:832–837.
- Ash ASF and Schild HO (1966) Receptors mediating some actions of histamine. *Br J Pharmacol* 27:427–439.
- Barger G and Dale HH (1910) Chemical structure and sympathomimetic action of amines. *J Physiol (London)* 41:19–59.
- Baudry M, Martres MP and Schwartz JC (1975) H₁ and H₂ receptors in the histamine-induced accumulation of cyclic AMP in guinea pig brain slices. *Nature (London)* 253:362–363.
- Bertaccini G and Coruzzi G (1995) An update on histamine H₃ receptors and gastrointestinal functions. *Dig Dis Sci* 40:2052–2063.
- Black JW, Duncan WAM, Durant CJ, Ganellin CR and Parsons EM (1972) Definition and antagonism of histamine H₂-receptors. *Nature (London)* 236:385–390.
- Cheng Y-C and Prusoff WH (1973) Relation between the inhibition constant (K_i) and the concentration of inhibitor which causes fifty per cent inhibition (IC_{50}) of an enzymic reaction. *Biochem Pharmacol* 22:3099–3108.
- Clark MA, Korte A and Egan RW (1993) Guanine nucleotides and pertussis toxin reduce the affinity of histamine H₃ receptors on AtT-20 cells. *Agents Actions* 40:129–134.
- Gantz I, Schaffer M, DelValle J, Logsdon C, Campbell V, Uhler M and Yamada T (1991) Molecular cloning of a gene encoding the histamine H₂ receptor. *Proc Natl Acad Sci USA* 88:429–433.
- Hill SJ, Ganellin CR, Timmerman H, Schwartz JC, Shankley NP, Young JM, Schunack W, Levi R and Haas HL (1997) International Union of Pharmacology. XIII. Classification of histamine receptors. *Pharmacol Rev* 49:253–278.
- Hinuma S, Habata Y, Fujii R, Kawamata Y, Hosoya M, Fukusumi S, Kitada C, Masuo Y, Asano T, Matsumoto H, Sekiguchi M, Kurokawa T, Nishimura O, Onda H and Fujino M (1998) A prolactin-releasing peptide in the brain. *Nature (London)* 393:272–276.
- Imamura M, Seyedi N, Lander HM and Levi R (1995) Functional identification of histamine H₃-receptors in the human heart. *Circ Res* 77:206–210.
- Kathmann M, Schlicker E and Goethert M (1994) Intermediate affinity and potency of clozapine and low affinity of other neuroleptics and of antidepressants at H₃ receptors. *Psychopharmacology* 116:464–468.
- Laitinen JT and Jokinen M (1998) Guanosine 5'-(γ -[³⁵S]thio) triphosphate autoradiography allows selective detection of histamine H₃ receptor-dependent G protein activation in rat brain tissue sections. *J Neurochem* 71:808–816.
- Leurs R, Blandina P, Tedford C and Timmerman H (1998) Therapeutic potential of histamine H₃ receptor agonists and antagonists. *Trends Pharmacol Sci* 19:177–183.
- Leurs R, Kathmann M, Vollinga RC, Menge WMPB, Schlicker E and Timmerman H (1996) Histamine homologs discriminating between two functional H₃ receptor assays. Evidence for H₃ receptor heterogeneity? *J Pharmacol Exp Ther* 276:1009–1015.
- Molderings GJ, Weissenborn G, Schlicker E, Likungu J and Goethert M (1992) Inhibition of noradrenaline release from the sympathetic nerves of the human saphenous vein by presynaptic histamine H₃ receptors. *Naunyn-Schmiedeberg's Arch Pharmacol* 346:46–50.
- Raible DG, Lenahan T, Fayvilevich Y, Kosinski R and Schulman ES (1994) Pharmacologic characterization of a novel histamine receptor on human eosinophils. *Am J Respir Crit Care Med* 149:1506–1511.
- Reinscheid RK, Nothacker H-P, Bourson A, Ardati A, Henningsen RA, Bunzow JR, Grady DK, Langen H, Monsma FJ Jr and Civelli O (1995) Orphanin FQ: A neuropeptide that activates an opioidlike G protein-coupled receptor. *Science (Wash DC)* 270:792–794.
- Rodriguez AA, Jansen FP, Leurs R, Timmerman H and Prell GD (1995) Interaction of clozapine with the histamine H₃ receptor in rat brain. *Br J Pharmacol* 114:1523–1524.
- Sakurai T, Amemiya A, Ishii M, Matsuzaki I, Chemelli RM, Tanaka H, Williams SC, Richardson JA, Kozlowski GP, Wilson S, Arch JRS, Buckingham RE, Haynes AC, Carr SA, Annan RS, McNulty DE, Liu W-S, Terrett JA, Elshourbagy NA, Bergsma DJ and Yanagisawa M (1998) Orexins and orexin receptors: A family of hypothalamic neuropeptides and G protein-coupled receptors that regulate feeding behavior. *Cell* 92:573–585.
- Schlicker E, Fink K, Hinterthaler M and Goethert M (1989) Inhibition of noradrenaline release in the rat brain cortex via presynaptic H₃ receptors. *Naunyn-Schmiedeberg's Arch Pharmacol* 340:633–638.
- Schlicker E, Kathmann M, Bitschnau H, Marr I, Reidemeister S, Stark H and Schunack W (1996) Potencies of antagonists chemically related to iodoproxyfan at histamine H₃ receptors in mouse brain cortex and guinea pig ileum: Evidence for H₃ receptor heterogeneity? *Naunyn-Schmiedeberg's Arch Pharmacol* 353:482–488.
- Schwartz JC, Lampart C, Rose C, Rehault MC, Bischoff S and Pollard H (1970) Development of the histaminergic systems in the neonatal rat brain. *J Physiol (Paris)*, 62(Suppl):447.
- Simmons DM, Arriza JL and Swanson LW (1989) A complete protocol for in situ hybridization of messenger RNAs in brain and other tissues with radiolabeled single-stranded RNA probes. *J Histochem* 12:169–181.
- Smits RPJM and Mulder AH (1991) Inhibitory effects of histamine on the release of serotonin and noradrenaline from rat brain slices. *Neurochem Int* 18:215–220.
- Stark H, Purand K, Huels A, Ligneau X, Garbarg M, Schwartz J-C and Schunack W (1996a) [125I]Iodoproxyfan and related compounds: A reversible radioligand and novel classes of antagonists with high affinity and selectivity for the histamine H₃ receptor. *J Med Chem* 39:1220–1226.
- Stark H, Schlicker E and Schunack W (1996b) Developments of histamine H₃-receptor antagonists. *Drugs Future* 21:507–520.
- Tatemoto K, Hosoya M, Habata Y, Fujii R, Kakegawa T, Zou M-X, Kawamata Y, Fukusumi S, Hinuma S, Kitada C, Kurokawa T, Onda H and Fujino M (1998) Isolation and characterization of a novel endogenous peptide ligand for the human APJ receptor. *Biochem Biophys Res Commun* 251:471–476.
- West RE Jr, Zweig A, Shih NY, Siegel MI, Egan RW and Clark MA (1990) Identification of two H₃-histamine receptor subtypes. *Mol Pharmacol* 38:610–613.
- Yamashita M, Fukui H, Sugama K, Horio Y, Ito S, Mizuguchi H and Wada H (1991) Expression cloning of a cDNA encoding the bovine histamine H₁ receptor. *Proc Natl Acad Sci USA* 88:11515–11519.

Send reprint requests to: Dr. Timothy W. Lovenberg, R.W. Johnson Pharmaceutical Research Institute, 3535 General Atomics Ct., San Diego, CA. E-mail: tlovenbe@prius.jnj.com

Orphan G protein-coupled receptors: a neglected opportunity for pioneer drug discovery

Jeffrey M. Stadel, Shelagh Wilson and
Derk J. Bergsma

Access to DNA databases has introduced an exciting new dimension to the way biomedical research is conducted. 'Genomic research' offers tremendous opportunity for accelerating the identification of the cause of disease at the molecular level and thereby foster the discovery of more selective medicines to improve human health and longevity. The current challenge is to close the gap rapidly between gene identification and clinical development of efficacious therapeutics. In the present review, **Jeffrey Stadel, Shelagh Wilson and Derk Bergsma** outline the rationale and describe strategies for converting one large class of novel genes, orphan G protein-coupled receptors (GPCRs), into therapeutic targets. Historically, the superfamily of GPCRs has proven to be among the most successful drug targets and consequently these newly isolated orphan receptors have great potential for pioneer drug discovery.

The advent of rapid DNA sequencing spawned the 'genomic era', which has led to the initiation of the Human Genome Project. The novel technologies developed in association with genomic research have already had a significant impact on the way investigations into the basis of disease are being conducted and will, no doubt, substantially enhance the means by which diseases are diagnosed and treated in the near future. To keep pace with the evolution of molecular medicine, the pharmaceutical industry has embraced genomics and is attempting to exploit the new technologies to identify novel targets for drug discovery. The major questions that remain to be addressed concern how to convert genomic sequences into therapeutic targets in an expeditious manner and eventually to obtain pharmaceutical drugs that will enhance the quality of life. This review will deal with a single class of novel molecular targets, focusing on the burgeoning collection of G protein-coupled receptors (GPCRs) called 'orphan' receptors¹. GPCRs are a superfamily of integral plasma membrane proteins involved in a broad array of signalling pathways. Since the first cloning of GPCR gene sequences over a decade ago, novel members of the GPCR

superfamily have continued to emerge through cloning activities as well as through bioinformatic analyses of sequence databases, although their ligands are unidentified and their physiological relevance remain to be defined. These 'orphan' receptors provide a rich source of potential targets for drug discovery.

The members of the GPCR superfamily are related both structurally and functionally. The signature motif of these receptors is seven distinct hydrophobic domains, each of which is 20–30 amino acids long and which are linked by hydrophilic amino acid sequences of varied length^{2,3}. Biophysical⁴ and biochemical⁵ studies support the notion that these receptors are intercalated into the plasma membrane with the amino terminus extracellular and the carboxy terminus in the cytoplasmic portion of the cell. Therefore, these receptors are often referred to as seven transmembrane (or 7TM) receptors. While it is not yet known how many individual genes actually encode these receptors, it is clear that this family of proteins is one of the largest yet identified. Functionally, GPCRs share in common the property that upon agonist binding they transmit signals across the plasma membrane through an interaction with heterotrimeric G proteins^{6,7}. These receptors respond to a vast range of agents^{2,5,8} such as protein hormones, chemokines, peptides, small biogenic amines, lipid-derived messengers, divalent cations (e.g. a Ca^{2+} sensor has been identified that is a GPCR)⁹ and even proteases such as thrombin, which activates its receptor by cleaving off a portion of the amino terminus¹⁰. Finally, these receptors play an important role in sensory perception including vision and smell^{2,5,8}. Correlated with the broad range of agents that activate these receptors is their existence in a wide variety of cells and tissue types, indicating that they play roles in a diverse range of physiological processes. It is likely, therefore, that the GPCR superfamily is involved in a variety of pathologies. This point was recently emphasized by the surprising discovery that certain GPCRs for chemokines act as co-factors for HIV infection^{11–13}.

GPCRs represent the primary mechanism by which cells sense alterations in their external environment and convey that information to the cells' interior. The binding of an agonist to the receptor promotes conformational changes in the cytoplasmic domains that lead to the interaction of the receptor with its cognate G protein(s). Agonist-promoted coupling between receptors and G proteins leads to the activation of intracellular effectors that substantially amplify the production of second messengers feeding into the signalling cascade. Since effectors are often enzymes [e.g. adenylate cyclase¹⁴, which converts ATP to cAMP, or phospholipase C (Ref. 15), which hydrolyses inositol lipids in membranes to release inositol trisphosphate, which in turn mobilizes Ca^{2+} within a cell] or ion channels¹⁶, many second messenger molecules can be produced as the result of a single agonist binding event with its receptor. Changes in the intracellular levels of ions or cAMP, or both,

J. M. Stadel,
Associate Director,
Department of
Cardiovascular
Pharmacology,
SmithKline Beecham
Pharmaceuticals, 709
Swedeland Road,
King of Prussia,
PA 19406, USA.
S. Wilson,
Assistant Director,
Department of
Molecular Screening
Technologies, New
Frontiers Scientific
Park (North), Third
Avenue, Harlow,
UK CM19 5AW,
and
D. J. Bergsma,
Director,
Department of
Molecular Genetics,
SmithKline Beecham
Pharmaceuticals, 709
Swedeland Road,
King of Prussia,
PA 19406, USA.

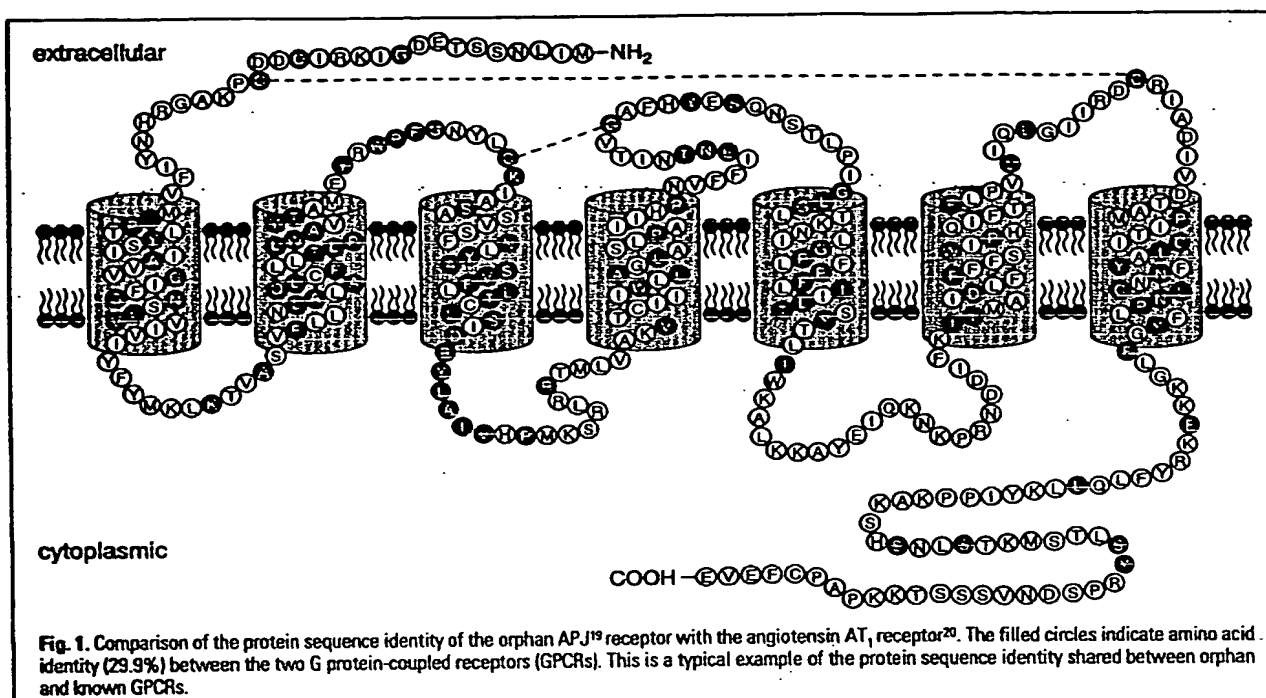


Fig. 1. Comparison of the protein sequence identity of the orphan APJ¹⁹ receptor with the angiotensin AT₁ receptor²⁰. The filled circles indicate amino acid identity (29.9%) between the two G protein-coupled receptors (GPCRs). This is a typical example of the protein sequence identity shared between orphan and known GPCRs.

result in the modulation of distinct phosphorylation cascades^{17,18}, extending through the cytosol to the nucleus, that eventually culminate in the physiological response of the cell to the extracellular stimulus. Although the overall paradigm is apparently the same for all GPCRs, the diversity of receptors, G proteins and effectors suggest a myriad of potential signalling processes and this becomes an important concept as we try to identify the function of orphan GPCRs.

To date, more than 800 GPCRs have actually been cloned from a variety of eukaryotic species, from fungi to humans [see L. F. Kolakowski in GCRDb-WWW The G Protein-Coupled Receptor DataBase World-Wide-Web Site (<http://receptor.mgh.harvard.edu/GCRDBHOME.html>)]. For humans, the most represented species, about 140 GPCRs have been cloned for which the cognate ligands are also known. This number excludes the sensory olfactory receptors, of which hundreds to thousands are predicted to exist. By traditional molecular genetic approaches, coupled with the explosion in genomic information, it has been possible to identify more than 100 additional orphan GPCR family members. By definition, there is enough sequence information in the receptor cDNAs to place them clearly in the superfamily of GPCRs, but often there is insufficient sequence homology with known members of this family to be able to assign their ligands with confidence or predict their function. In total, there are currently over 240 human GPCRs, excluding sensory receptors. As the size of sequence databases continues to increase, this list is expected to grow to 400, and perhaps even to 1000 or more unique gene products. The list will grow even further as paralogues and alternatively spliced GPCR variants emerge. Most orphan GPCRs share a low degree of

sequence homology (typically about 25–35% overall amino acid sequence identity), with known GPCRs, suggesting that they belong to new subgroups of receptors (Fig. 1)^{19,20}. Indeed, several orphan GPCRs show closer homology to each other than to known GPCRs. Nevertheless, the majority of orphan receptors are phylogenetically distributed among a broad spectrum of distantly related, known receptor subgroups.

What is the rationale for investing considerable time and resources into trying to establish the function of orphan GPCRs? Simply stated, GPCRs have a proven history of being excellent therapeutic targets. Within the past 20 years, several hundred new drugs have been registered that are directed towards activating or antagonizing GPCRs; in fact, it is estimated that most current research within the pharmaceutical industry is focused on this signalling pathway²¹. Table 1 shows a representative snapshot of a variety of receptors, disease targets and corresponding drugs. It is clear from this table that the therapeutic targets span a wide range of disorders and disease states. Another example of the significance and versatility of GPCRs is the number of cases of genetic diseases that are linked to defects in these proteins; some of these diseases are indicated in Table 2 (Refs 22–38). It is likely that many more genetic diseases will be mapped to GPCRs as the era of genomics continues to expand and families with inherited mutations are examined much more comprehensively.

The importance of GPCRs to drug discovery continues to be manifested by the fact that across the pharmaceutical industry active research projects, ranging from basic studies all the way through to advanced development, are focused on GPCRs as primary targets. Molecular biology has had a dramatic influence on these efforts.

Table 1. Examples of marketed drugs for G protein-coupled receptors (GPCRs)

GPCR	Generic	Drug	Indication
Muscarinic acetylcholine	Bethanechol	Urecholine	GI <i>urinary retention</i>
	Dicyclomine	Bentyl	GI
	Ipratropium	Atrovent	CP <i>pulmonary disease</i>
Adrenoceptor			
β_1	Atenolol	Tenormin	CP <i>hypertension</i>
α_2	Clonidine	Catapres	CP <i>"</i>
β_1/β_2	Propranolol	Inderal	CP <i>"</i>
α_1	Terazosin	Hytrin	CP <i>"</i>
β_2	Albuterol	Ventolin	CP <i>asthma</i>
$\beta_1/\beta_2/\alpha_1$	Carvedilol	Coreg	CP <i>hypertension, congestive heart failure</i>
Angiotensin			
AT ₁	Losartan	Cozaar	CP
	Eprosartan	Teveten	CP
Calcitonin	Calcitonin	Calcimar	Osteoporosis
	eel-Calcitonin	Elcatonin	Osteoporosis
Dopamine			
D ₂	Metoclopramide	Reglan	GI
D ₂ /D ₃	Ropinirole	Requip	CNS
D ₂	Haloperidol	Haldol	CNS
Gonadatropin-releasing factor	Goserelin	Zoladex	Cancer
	Nafarelin	Synarel	Endometriosis
Histamine			
H ₁	Dimenhydrinate	Dramamine	CNS
H ₁	Terfenadine	Seldane	CP
H ₂	Cimetidine	Tagamet	GI
H ₂	Ranitidine	Zantac	GI
Serotonin (5-HT)			
5-HT _{1D}	Sumatriptan	Imitrex	CNS <i>migraine</i>
5-HT _{2A}	Ritanserin	Tisertan	CNS
5-HT ₄	Cisapride	Propulsid	GI <i>motility</i>
5-HT _{1B}	Trazodone	Desyrel	CNS <i>depression</i>
5-HT _{2A/2C}	Clozapine	Clozaril	CNS <i>schizophrenia</i>
Leukotriene	Pranlukast	Onon	CP
	Zafirlukast	Accolate	CP
Opioid			
κ	Buprenorphine	Buprenex	CNS
	Butorphanol	Stadol	CNS
μ	Alfentanil	Alfenta	CNS
	Morphine	Kadian	CNS
Oxytocin		Syntocinon	Labour
Prostaglandin	Epoprostenol	Flolan	CP
	Misoprostol	Cytotec	GI
Somatostatin	Octreotide	Sandostatin	Cancer
Vasopressin	Desmopressin		CP/Renal

CP, cardiopulmonary system; GI, gastrointestinal system.

Table 2. Diseases associated with mutations of G protein-coupled receptors (GPCRs)

GPCR	Mutation	Disease	Refs
Rhodopsin	Missense: Pro23 to His (NT) Missense: Val87 to Asp (2TM) Missense: Tyr178 to Cys (2EL) Nonsense: Gln344 to Stop (CT)	Retinitis pigmentosa	22, 23
Thyroid stimulating hormone	Missense: Asp619 to Gly (3IL) Missense: Ala623 to Ile (3IL)	Hyperfunctioning thyroid adenomas	24
Luteinizing hormone	Missense: Asp578 to Gly (6TM)	Precocious puberty	25
Vasopressin V ₂	Missense: Arg137 to His (2IL) Missense: Gly185 to Cys (2EL) Frameshift at Arg230 (3TM)	X-linked nephrogenic diabetes	26–28
Ca ²⁺	Missense: Arg186 to Glu (NT) Missense: Glu298 to Lys (NT) Missense: Arg796 to Trp (3IL) Missense: Glu128 to Ala (NT)	Hyperparathyroidism, hypocalciuric hypercalcaemia	29, 30
Parathyroid hormone (PTH type b)	Missense: His223 to Arg (1IL)	Short-limbed dwarfism	31
β ₃ -Adrenoceptor	Missense: Trp64 to Arg (1IL)	Obesity, NIDDM	32–34
Growth-hormone-releasing hormone	Nonsense: Glu72 to Stop (NT)	Dwarfism	35
Adrenocorticotropin	Missense: Ser74 to Ile (2TM)	Glucocorticoid deficiency	36
Glucagon	Missense: Gly40 to Ser (NT)	Diabetes, hypertension	37, 38

Abbreviations: CT, carboxyl terminus; EL, extracellular loop; IL, intracellular loop; NIDDM, non-insulin-dependent diabetes mellitus; NT, amino terminus; TM, transmembrane segment.

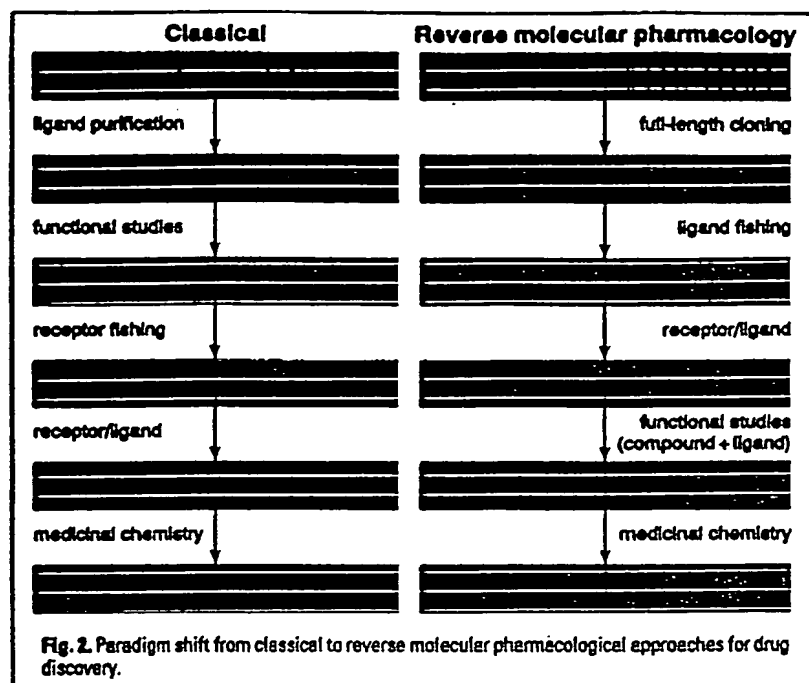
The cloning of cDNAs for well-known GPCRs led to the discovery of a surprising number of paralogues⁵. The existence of these novel receptor subtypes was unexpected because the current cornucopia of pharmacological agents does not possess the required selectivity to distinguish all of them clearly, and thus an opportunity for drug discovery was quickly recognized. Current research efforts seek to define the physiology associated with these novel receptor subtypes and to discover highly selective compounds as potential pharmaceutical drugs. These efforts are almost exclusively focused on GPCRs for which activating ligands are known. Since characterized GPCRs were, and continue to be, attractive therapeutic targets, it is most reasonable to speculate that many of the orphan receptors have similar potential. The initial challenge is to determine the function of each orphan receptor through the identification of activating ligands and, once the function is clarified, link the orphan receptor to a specific disease and thus establish it as a candidate for a comprehensive drug discovery effort.

Reverse molecular pharmacology

Until recently, research into the identification of GPCRs as targets for drug discovery has been conducted using the traditional approach illustrated in Fig. 2. For this strategy, the starting point is functional activity, which forms the basis of an assay by which a ligand is

identified through purification from biological fluids, cell supernatants or tissue extracts. One example of the success of this strategy is the discovery of the potent vasoconstricting peptide endothelin³⁹. Once isolated, the ligand is used to characterize its cellular and tissue biology as well as its pathophysiological role. Subsequently, cDNAs encoding corresponding receptors are 'fished' from gene libraries using a variety of methodologies (e.g. receptor purification and expression cloning) that often either directly or indirectly use the ligand as the 'hook'. As the nucleotide sequences for GPCRs begin to accumulate and be analysed, additional receptors can be cloned by homology screening, by positional cloning, and by polymerase chain reaction (PCR) methodologies that use oligonucleotide primers based on nucleotide sequences conserved within the seven transmembrane domains of the GPCR family. Once the cloned human receptor cDNA is expressed in a heterologous cell system⁴⁰, it is used, together with its ligand, to form the basis of a screen to explore chemical compound libraries for receptor antagonists or agonists. Lead structures identified in the screen are refined through medicinal chemistry using an iterative process. Resulting drug leads with appropriate *in vivo* pharmacology are passed on into the clinic for development.

Recently, this paradigm has changed radically with the introduction of a new reverse molecular pharmacological



strategy, shown diagrammatically in Fig. 2. Through both traditional molecular cloning techniques and, more recently, mass sequencing of expressed sequence tags (ESTs) from cDNA libraries, it is now possible to identify GPCRs through computational or bioinformatic methodologies. The EST approach, initially proposed by Sidney Brenner (University of Cambridge) and first brought to large-scale practice by Craig Venter (The Institute of Genome Research), constitutes random, single-pass sequencing of cDNAs randomly picked from a collection of cDNA libraries, followed by extensive bioinformatic analysis of the sequence to identify structural signatures characteristic of GPCRs. Once new members of the GPCR superfamily are identified, the recombinantly expressed receptors are used in functional assays to search for the associated novel ligands. The receptor–ligand pair are then used for compound bank screening to identify a lead compound that, together with the activating ligand, is used for biological and pathophysiological studies to determine the function and potential therapeutic value of a receptor antagonist (or agonist) in ameliorating a disease process. In addition, clues as to therapeutic potential may involve receptor genotyping of disease populations. Once a link with a disease is finally identified, an appropriate compound can be advanced for clinical study.

The reverse molecular pharmacological strategy is a far more daunting challenge and risky endeavour when compared with the more traditional approach, since the starting material for a drug discovery effort is simply an orphan receptor of unknown function, with no apparent relationship to a disease indication. However, the potential reward of using this approach is that resultant drugs naturally will be pioneer or innovative discoveries, and a

significant proportion of these unique drugs may be useful to treat diseases for which existing therapies are lacking or insufficient.

Screening strategy

Figure 3 illustrates the generic strategy that we use for our reverse molecular pharmacological approach. In addition to the EST approach, which has yielded the majority of our collection of orphan receptors, we have also used a number of more traditional approaches such as low-stringency screening, using portions of known GPCRs as hybridization probes, as well as PCR-based methods. By these techniques we have succeeded in identifying more than 70 orphan receptors in addition to those already cited in the literature.

Since cDNAs identified by EST cloning are often incomplete, northern hybridization analysis is used to establish the tissue or cell pattern of mRNA expression of the GPCRs. This information is used to identify the tissue or cell cDNA libraries that are to be probed for full-length clones and, significantly, to determine whether a receptor is expressed in a particular normal or diseased tissue of interest. A highly selective tissue expression pattern may also provide a clue with respect to receptor function. Once obtained, full-length GPCR clones are expressed in mammalian cell lines and yeast model systems (see below) for functional analysis. *Xenopus* oocytes may also be used for expression; however, low screening throughput limits their use to a secondary, confirmatory assay system. For mammalian cell expression, the human embryonic kidney (HEK) 293 cell line or Chinese hamster ovary (CHO) cells are frequently used. These cell types possess a large repertoire of G proteins that are necessary for coupling to downstream effectors *in situ*. They also share a reliable history of positive functional coupling for a wide variety of known GPCRs. However, since receptor coupling cannot be accurately predicted from primary sequence data, orphan GPCRs may need to be expressed in a variety of cell lines to establish viable coupling.

These heterologous expression systems form the basis for screening for an activating ligand. The success of establishing functional coupling of the recombinant receptor depends to a large extent on whether the receptor is properly expressed, which may be assessed by northern or Western blot analysis, and whether appropriate G proteins and downstream effectors are present in the cell in which the receptor is expressed. There are several major technical challenges to be met in order to initiate ligand fishing. Because it is difficult to predict accurately the coupling specificity of orphan GPCRs from their primary sequence, assays must be chosen that will detect a wide range of coupling mechanisms. These generally focus on changes in intracellular levels of cAMP or Ca^{2+} but can also include more generic measurements, such as metabolic activation of the cell via the cytosensor microphysiometer⁴¹. Recently, it has become possible to implement most of these screens in high-throughput format by using fluorescent-based

assays and using charge-coupled device cameras and reporter gene constructs that allow easy readout of the assay on microtitre plates. Ever increasing throughput of the assays will be necessary to screen large libraries. However, this approach is somewhat cumbersome and inefficient if all the assays described above have to be used. Is it possible to funnel heterologous signal transduction through a defined pathway? The prospect of an assay for a single transduction pathway comes from the observation that heterologous expression of the G protein subunit $G_{\alpha 15/16}$ promoted coupling of various GPCR subfamily members through activation of phospholipase $C\beta$ and likely Ca^{2+} mobilization^{42,43}. Although this approach may not work universally, the diversity of the GPCRs successfully coupled through $G_{\alpha 16}$ to phospholipid metabolism suggests that this could be a useful method to screen for orphan receptor activation.

Once heterologous receptor expression is achieved and functional assays are in place, ligand fishing experiments can be initiated. Although the homology with known GPCRs is low, we nevertheless begin by screening the orphans against known GPCR ligands; since the sequence homology between some subtypes of known receptors can be low (e.g. 30–40% between neuropeptide Y receptor subtypes), it is possible that new paralogue receptors for known ligands still remain to be discovered. The next step is to search for novel activating ligands by screening biological extracts obtained from tissues, biological fluids and cell supernatants. An additional option is screening libraries of compounds for activating ligands. Complex libraries of peptides or compound collections could be rich sources of 'surrogate' agonists that would promote receptor activation and coupling but are not endogenous ligands. The rationale for searching for surrogate agonists springs from a report that a nonpeptide agonist has been discovered for the angiotensin II receptor⁴⁴. There is also an obvious precedent for nonpeptide agonists for opioid receptors. Screening of the very large libraries that will be generated by fractionation of biological extracts and by combinatorial chemical synthesis requires that the functional assays used have not only a high throughput but are also robust, since false positives can be a significant problem.

Examples are beginning to emerge from several efforts showing that progress has been made in characterizing orphan GPCRs. A first example is the identification of an orphan GPCR that functions as a calcitonin gene-related peptide (CGRP) receptor⁴⁵. CGRP is a peptide of 37 amino acids, widely distributed in neurones, and functions as a potent vasodilator. It may be involved in migraine and has been implicated in non-insulin-dependent diabetes mellitus because it promotes resistance to insulin. An orphan GPCR EST was derived from a human synovium cDNA library⁴⁵. Sequence analysis showed that the new GPCR has ~56% similarity to the human calcitonin receptor and was hence originally expected to be a new subtype of the calcitonin receptor. The message for this novel receptor was expressed

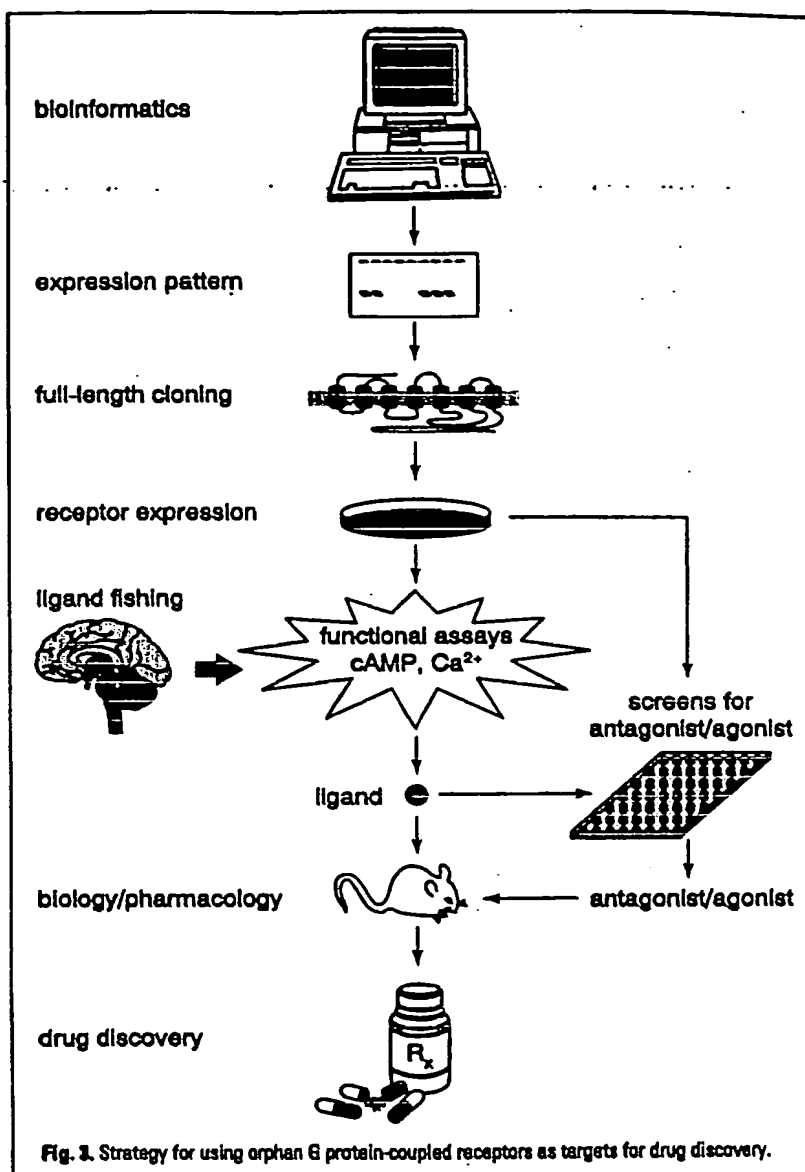


Fig. 3. Strategy for using orphan G protein-coupled receptors as targets for drug discovery.

predominantly in lung, which is known to be a relatively rich source of CGRP receptors. Following full-length cloning from a human lung library, the orphan receptor cDNA was stably expressed in HEK293 cells. Both radioligand binding using ^{125}I CGRP, as well as functional assays of CGRP-stimulated cAMP accumulation, demonstrated an appropriate pharmacological profile for the expressed receptor similar to that observed with endogenous CGRP receptors on human neuroblastoma cells. In addition to identifying the CGRP receptor, the reverse molecular pharmacology approach has also been used to identify other orphan receptors, such as the anaphylatoxin C3a receptor⁴⁶.

The examples given above are for receptors with significant homology to known GPCR superfamily members and their activating ligands proved to be known GPCR ligands. Will ligand fishing be successful in identifying novel endogenous ligands? Recently, two groups

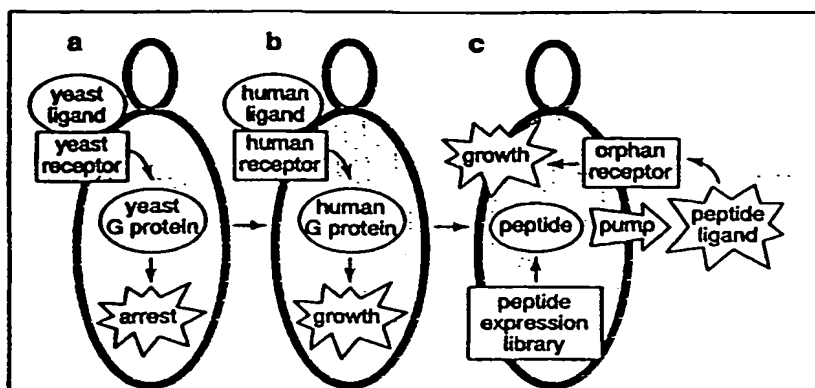


Fig. 4. Yeast-based screen for the identification of agonists for orphan G protein-coupled receptors (GPCRs). **a:** Normal, endogenous GPCR signalling in yeast (*Saccharomyces cerevisiae*). **b:** Substitution of a human GPCR and a human G_{α} subunit for yeast counterparts and modification of downstream signalling pathways such that agonist stimulation of the recombinant GPCR promotes growth. This yeast strain can be screened using biological extracts or compound libraries, or both. **c:** Yeast cells can be engineered to secrete small peptides from a random peptide library to identify autocrine surrogate, peptide agonists for recombinant orphan GPCRs. Modified from Ref. 49, with permission.

investigated an orphan opioid-like receptor, ORL1 (Refs 47 and 48). Both groups expressed the orphan GPCR in CHO cells and challenged the transfected cells with a series of opiate agonists, but without response. Both groups then used a similar ligand fishing approach. Taking crude extracts from rat brain⁴⁷ or porcine brain⁴⁸, they screened against the stably transfected cell lines using inhibition of adenylate cyclase activity as a functional assay. They were able to fractionate the brain extracts and identify the novel dynorphin-like ligand, which they called nociceptin⁴⁷ or orphanin FQ (Ref. 48). Thus, both teams successfully established a functional assay in transfected CHO cells that allowed the purification of a novel neuropeptide ligand that is 17 amino acids long for the orphan receptor. This work validates the ligand fishing approach for characterizing the function of orphan GPCRs.

Concluding remarks and future challenges

Although orphan GPCRs have been around for over ten years, very few companies have, until recently, been willing to risk their resources to explore opportunities among this category of receptors. However, the environment for the pharmaceutical industry has changed due to the confluence of several major technological advances. The conversion of gene sequences encoding GPCRs to drug targets is substantially aided by the development of combinatorial chemistry methods and miniaturized high-throughput screening techniques. The future challenge for drug discovery in this arena is to integrate these technologies innovatively and productively. One glimpse of the future comes from the field of functional genomics. The endogenous GPCR transduction system of the yeast, *Saccharomyces cerevisiae*, which is the pheromone pathway required for conjugation and mating, has been commandeered – through genetic engineering – to permit functional expression and coupling of human GPCRs and

humanized G protein subunits to the endogenous signalling machinery^{49–51} (Fig. 4). Further manipulations involve conversion of the normal yeast response to pheromone or activating ligand (growth arrest) to positive growth on selective media or to reporter gene expression. In addition, yeast cells have been engineered to express and secrete small peptides from a random peptide library that will permit the autocrine activation of heterologously expressed human GPCRs (Refs 49 and 51). This provides an elegant means of screening rapidly for surrogate peptide agonists that activate orphan receptors. This yeast system is, of course, not limited to autocrine ligand screening but can also be used in high-throughput mode to screen directly the fractions from biological extracts and the various chemical libraries as described above. A major advantage of the yeast system over the mammalian heterologous expression systems is its ease of use and its lack of endogenous GPCRs, which can confound ligand fishing expeditions in mammalian cells.

There is now tremendous pressure to be the first on the market with highly selective drugs that target therapeutic areas of unmet medical need and ideally have novel mechanisms of action. As a consequence, the pharmaceutical industry has recognized the power of genomics to provide it with new and unique drug targets. Genomics has responded with a plethora of novel proteins, included among them over 100 orphan GPCRs. Because of the proven link of GPCRs to a wide variety of diseases and the historical success of drugs that target GPCRs, we believe that these orphan receptors are among the best targets of the genomic era to advance into the drug discovery process.

Selected references

- Libert, F., Vassart, G. and Parmentier, M. (1991) *Curr. Opin. Cell Biol.* 3, 218–223
- Strader, C. D., Fong, T. M., Tota, M. R. and Underwood, D. (1994) *Annu. Rev. Biochem.* 63, 101–132
- Baldwin, J. M. (1994) *Curr. Opin. Cell Biol.* 6, 180–190
- Schertler, G. F. X., Villa, C. and Henderson, R. (1993) *Nature* 362, 770–772
- Dohlman, H. G., Thorner, J., Caron, M. G. and Lefkowitz, R. J. (1991) *Annu. Rev. Biochem.* 60, 653–688
- Neer, E. J. (1995) *Cell* 80, 249–257
- Rens-Domiano, S. and Hamm, H. E. (1995) *FASEB J.* 9, 1059–1066
- Coughlin, S. R. (1994) *Curr. Opin. Cell Biol.* 6, 191–197
- Brown, E. M. et al. (1993) *Nature* 366, 575–580
- Vu, T.-K. H., Hung, D. T., Wheaton, V. I. and Coughlin, S. R. (1991) *Cell* 64, 1057–1068
- Feng, Y., Broder, C. C., Kennedy, P. E. and Berger, E. A. (1996) *Science* 272, 872–876
- Deng, H. K. et al. (1996) *Nature* 381, 661–666
- Dragic, T. et al. (1996) *Nature* 381, 667–673
- Sunahara, R. K., Dessauer, C. W. and Gilman, A. G. (1996) *Annu. Rev. Pharmacol. Toxicol.* 36, 461–480
- Rhee, S. G. and Choi, K. D. (1992) *Adv. Sec. Mess. Phosph. Res.* 26, 35–49
- Clapham, D. E. (1995) *Cell* 80, 259–268
- Hunter, T. (1995) *Cell* 80, 225–236
- Graves, J. D., Campbell, J. S. and Krebs, E. G. (1995) *Ann. New York Acad. Sci.* 766, 320–341
- O'Dowd, B. F. et al. (1993) *Gene* 136, 355–360
- Bergsma, D. J. et al. (1992) *Biochem. Biophys. Res. Commun.* 183, 989–995
- Roush, W. (1996) *Science* 271, 1056–1058
- Dryja, T. P. et al. (1990) *Nature* 343, 364–366
- Sung, C.-H. et al. (1991) *Proc. Natl. Acad. Sci. U. S. A.* 88, 6481–6485
- Parma, J. et al. (1993) *Nature* 365, 649–651

- 25 Shenker, A., Laue, L., Kosugi, S., Merendino, J. J., Minegishi, T. and Cutler, G. B. (1993) *Nature* 365, 652-654
- 26 van den Ouweland, A. M. W. et al. (1992) *Nat. Genet.* 2, 99-102
- 27 Pan, Y., Metzberg, A., Das, S. and Gitschier, J. (1992) *Nat. Genet.* 2, 103-106
- 28 Rosenthal, W., Antaramian, A., Gilbert, S. and Birnbaumer, M. (1993) *J. Biol. Chem.* 268, 13030-13033
- 29 Pollak, M. R. et al. (1993) *Cell* 75, 1297-1303
- 30 Pollak, M. R. et al. (1994) *Nat. Genet.* 8, 303-307
- 31 Schipsni, E., Kruse, K. and Juppner, H. (1995) *Science* 268, 98-100
- 32 Walston, J. et al. (1995) *New Engl. J. Med.* 333, 343-347
- 33 Widen, E., Lehto, M., Kanninen, T., Walston, J., Shuldiner, A. R. and Groop, L. C. (1995) *New Engl. J. Med.* 333, 348-351
- 34 Clement, K. et al. (1995) *New Engl. J. Med.* 333, 352-354
- 35 Wajnrach, M. P., Gertner, J. M., Harbison, M. D., Chua, S. C. and Leibel, R. L. (1996) *Nat. Genet.* 12, 88-90
- 36 Clark, A. J. L., McLoughlin, L. and Grossman, A. (1993) *Lancet* 341, 461-462
- 37 Hager, J. et al. (1995) *Nat. Genet.* 9, 299-304
- 38 Chambers, S. M. and Morris, B. J. (1996) *Nat. Genet.* 12, 122
- 39 Yanagisawa, M. et al. (1988) *Nature* 332, 411-415
- 40 Tate, C. G. and Grisham, R. (1996) *Trends Biotechnol.* 14, 426-430
- 41 McConnell, H. M. et al. (1992) *Science* 257, 1906-1912
- 42 Offermanns, S. and Simon, M. (1995) *J. Biol. Chem.* 270, 15175-15180
- 43 Milligan, G., Marshall, F. and Rees, S. (1996) *Trends Pharmacol. Sci.* 17, 235-237
- 44 Perlman, S., Schambye, H. T., Rivero, R. A., Greenlee, W. J., Hjorth, S. A. and Schwartz, T. W. (1995) *J. Biol. Chem.* 270, 1493-1496
- 45 Aiyar, N. et al. (1996) *J. Biol. Chem.* 271, 11325-11329
- 46 Ames, R. S. et al. (1996) *J. Biol. Chem.* 271, 20231-20234
- 47 Meunier, J.-C. et al. (1995) *Nature* 377, 532-535
- 48 Reinscheid, R. K. et al. (1995) *Science* 270, 792-794
- 49 Broach, J. R. and Thorne, J. (1996) *Nature* 384 (Suppl.), 14-16
- 50 Price, L. A., Kajkowski, E. M., Hadcock, J. R., Ozenberger, B. A. and Pausch, M. H. (1995) *Mol. Cell. Biol.* 15, 6188-6195
- 51 Manfredi, J. P. et al. (1996) *Mol. Cell. Biol.* 16, 4700-4709

Acknowledgements
The authors wish to thank Drs Robert Ruffolo, Christine Debouck, Paul England and George Livi for their critical comments, as well as their continued encouragement and support.

CA₁A₂X-competitive inhibitors of farnesyltransferase as anti-cancer agents

Charles A. Omer and Nancy E. Kohl

For Ras oncoproteins to transform mammalian cells, they must be post-translationally farnesylated in a reaction catalysed by the enzyme farnesyl-protein transferase (FPTase). Inhibitors of FPTase have therefore been proposed as anti-cancer agents. In this review Charles Omer and Nancy Kohl discuss the development of FPTase inhibitors that are kinetically competitive with the protein substrate in the farnesylation reaction. These compounds are potent and selective inhibitors of the enzyme that block the tumorigenic phenotypes of *ras*-transformed cells and human tumour cells in cell culture and in animal models.

Since the identification of farnesyl-protein transferase (FPTase) activity in mammalian cells, there has been an intense effort to develop inhibitors of this housekeeping enzyme for use as potential, novel anti-cancer agents^{1,2}. This idea stems from the fact that several of the proteins that regulate mammalian cell proliferation require a post-translational modification catalysed by this enzyme for biological activity. Efforts over the past eight years have yielded potent, cell-active inhibitors of FPTase that demonstrate anti-proliferative activity in cell culture and in rodent models of cancer.

The focus of the FPTase inhibitor (FTI) studies has been inhibition of the transforming activity of the Ras

oncoproteins. Three *ras* genes, Ha-, N- and Ki-*ras*, encode four highly homologous, 21 kD proteins, Ha-, N-, Ki4A- and Ki4B-Ras (Ki4A- and Ki4B-Ras are encoded by splice variants of the Ki-*ras* gene)³. Ras functions to regulate the transduction of extracellular growth-promoting signals from membrane-bound receptor tyrosine kinases to intracellular growth-regulatory pathways. Typical of the low-molecular-weight G proteins, Ras is active when bound to GTP and inactive when bound to GDP. Cycling from the active to the inactive form is accomplished by the intrinsic GTPase activity of the protein. Mutations in Ras that abolish the GTPase activity result in constitutively active forms of the protein. Such oncogenically mutated forms of Ras, particularly Ki4B-Ras, are found in approximately 30% of many human cancers including 90% of pancreatic cancers and 50% of colon cancers^{4,5}.

Ras is synthesized as a biologically inactive, cytosolic protein that localizes to the inner surface of the plasma membrane where it acquires biological activity following a series of post-translational modifications (see Ref. 6 for review). The first and obligatory step in this series is the transfer of a 15-carbon isoprenoid, farnesyl, from farnesyl diphosphate (FPP) to the sulphur atom of the cysteine residue located four amino acids from the carboxyl terminus of the protein. This cysteine residue is part of the CA₁A₂X motif found in all FPTase protein substrates, where C is cysteine, A₁ and A₂ are usually aliphatic amino acids and X is usually serine, methionine, glutamine, alanine or cysteine. Following farnesylation, A₁A₂X is proteolytically cleaved and the now C-terminal farnesylcysteine is methylated. In the case of all of the Ras proteins except Ki4B-Ras, palmitate groups are then added to cysteine residues upstream of the farnesylated cysteine. The demonstration that farnesylation is essential for the transforming ability of the Ras oncoproteins⁷⁻¹⁰ has spurred the development of inhibitors of the enzyme that catalyses this reaction, FPTase, as anti-cancer agents.

FPTase is a ubiquitously expressed, cytosolic enzyme comprised of two subunits, a 45 kDa α subunit and a 48 kDa β subunit⁶. Cross-linking studies have shown

C. A. Omer,
Senior Research
Fellow,
and
N. E. Kohl,
Director,
Department of Cancer
Research, Merck
Research
Laboratories, West
Point, PA 19486, USA

The Use of Cloned Human Receptors for Drug Design

Paul R. Hartig

INTRODUCTION

Cloned human receptors are increasingly used by pharmacologists and medicinal chemists for the development of novel, site-specific drug therapies. This approach can be expected to produce many new drugs with improved efficacy and fewer side effects. The availability of sets of transfected human receptors makes it possible to target drugs to single human proteins from the inception of a drug design project. A critical requirement of this technology is that human genes must express a human pharmacology when expressed in the host cell line chosen for transfection. Studies on the human serotonin 5-HT₁ and 5-HT_{1D} 5-HT_{1A} receptors demonstrate that the gene sequence appears to determine the expressed receptor's pharmacological properties, with only a minor role played by the cellular environment in which the receptor is expressed. The cloning and characterization of the 5-HT_{1B} receptor also demonstrate another emerging principle of molecular pharmacology: the equivalent G protein-coupled receptor gene in different species can encode proteins with strikingly different pharmacological properties. Another important issue is the relationship between agonist and antagonist binding sites. Studies comparing agonist—1-(2,5-dimethoxy-4-methylphenyl)-2-aminopropane (DOM)—and antagonist binding sites of the human 5-HT₂ receptor demonstrate the strong differences exhibited by these two binding states of the same receptor protein. Finally, the fact that receptors of the G protein-coupled or 7TM (7 transmembrane) receptor superfamily exhibit many properties in common allows receptor homologues to be used to predict certain drug-binding properties. Examples from the serotonergic and adrenergic receptor families are presented.

A MOLECULAR PHARMACOLOGY

Cloned human receptors, conveniently expressed in transfected mammalian cell lines, now make it possible to approach drug design from a truly molecular perspective. In the past several years, the amino acid sequences of many receptor genes have been determined. Most of these cloned receptors are members of a closely related superfamily of genes known as the G protein-coupled receptors (or 7TM receptors), because of their characteristic single

subunit structure with 7TM-spanning segments. Similar cloning successes are now being reported for multisubunit ligand-gated ion channels, including the GABA_A receptor, the 5-HT₃ receptor, and the N-methyl-D-aspartate receptor. Molecular pharmacologists are beginning to integrate this new information into their views of physiological and pharmacological processes and are introducing cloned human receptors into the drug development process, often from the very start of a drug design effort.

Expression of cloned human receptors in mammalian cell lines allows pharmacologists to develop assay systems that individually express each receptor subtype important to a drug design project. Host cell lines can be chosen that are devoid of any related receptor sites, producing clean, unambiguous assay systems. These subtype-specific human receptor assays will allow medicinal chemists to design drugs with high affinity for the desired site of action and low affinity for those receptor subtypes that may induce side effects. In many cases, this design strategy can be expected to produce more potent medications with fewer side effects because of an improved molecular targeting of the drug. Even in cases where a blended drug possessing a spectrum of receptor activities may be the desired endpoint, pure human subtype assays provide a drug design team with an important advantage: unambiguous assays of the affinity of the drug candidate at each human receptor site of interest.

The second advantage of the cloned receptor approach is that low abundance sites, which have been difficult to study in tissue preparations (e.g., autoreceptors), can now be isolated and expressed at high density for use in ligand screening and basic science investigations. The third, and perhaps most important, advantage is the fact that many new receptor subtypes have been and continue to be discovered by receptor cloning efforts, which are providing a rapid increase in the number of potential drug target sites and can be expected to lead to several new medications in the future.

SPECIES DIFFERENCES IN RECEPTORS

The pharmacological binding properties of receptors are often similar for the same receptor subtype in different species, although many exceptions do occur. When species differences do exist, they dictate that great care be used in the choice of receptor assays for the purpose of human drug design. The best situation is found when human receptors can be used to screen for drug activity and selectivity. Cloned human receptors are making human receptor screening possible. These clones are commonly expressed in mammalian or sometimes in bacterial expression systems (Chapoi et al. 1990) for use in drug-binding assays. This leads to the important question of what influence the cell-line will have on the pharmacological properties of the human receptors that they express. For the purpose of drug screening, host cell lines must be chosen

so that cloned human receptors will properly reproduce the binding properties of native human tissues.

The influence of the cell-line host on the pharmacological binding properties of cloned human receptor subtypes has been studied in some detail. The binding properties of the serotonin 5-HT₂ receptor have been known to differ in different species, especially for certain ergot compounds. For example, mesulergine exhibits approximately thirtyfold higher affinity for rat cortical membranes than for comparable human tissue (Pazos et al. 1984). Transfection of a cDNA clone encoding the human 5-HT₂ receptor subtype (Hartig et al. 1990) into mouse fibroblast cells leads to expression of a serotonin receptor whose binding properties match that of human rather than rat cortical membranes (table 1). The largest drug-binding differences are seen for mesulergine, which binds to both the transfected human receptor and to human cortical membranes with an apparent affinity of approximately 150 nM, thirtyfold weaker than its affinity for the rat cortex 5-HT₂ receptor. This species difference also is seen for ritanserin, which belongs to an entirely different chemical class (table 1). In both cases, the transfected human receptor exhibits binding affinities in close agreement with human cortical tissue, even though the human gene has been expressed in a rodent, nonneuronal cell line. Together, these observations suggest that

TABLE 1. *Ligand-binding properties of a human 5-HT₂ receptor gene expressed in mouse fibroblast cells. A cDNA clone encoding the human 5-HT₂ receptor (Hartig et al. 1990) was expressed in mouse fibroblast L-M(tk-) cells. Membrane preparations were labeled with [³H]5-HT, and apparent dissociation constants were determined as described in Branchek and colleagues (1990). Comparative values from human cortex and rat cortex assays (Hoyer et al. 1986) are also provided.*

Drug	Apparent Dissociation Constant (K _d in nM) for Displacement of [³ H]Ritanserin Binding		
	Human Clone	Human Cortex	Rat Cortex
Ritanserin	1.1 ± 0.16	1.3	7.2
Cyproheptidine	2.9 ± 0.10	6.3	1.8
Mesulergine	129 ± 7.9	151	4.7
5-HT	598 ± 52	174	79
Olpazpine	2,111 ± 295	3,802	1,549
5-CT	7,790 ± 50	8,130	21,878

SOURCE: Part of the data in this table was previously described in Hartig and colleagues (1992).

it is the amino acid sequence of the receptor (nature), rather than the cellular environment in which the receptor gene is expressed and processed (nurture), that determines the species-specific pharmacological properties of the receptor. Further studies are needed to determine whether this will prove to be a general property of neurotransmitter receptors. For most cases examined so far, it appears that transfection of a human 7TM receptor gene into mammalian cell lines has produced ligand-binding properties in good agreement with previous binding assays in human brain tissue preparations. Furthermore, the use of different cell lines as transfection hosts was not found to substantially affect the pharmacological binding properties of a series of human G-protein-coupled receptors that the author and colleagues at Synaptic Pharmaceuticals have been investigating. This information provides a welcome degree of freedom in the choice of host cells for transfection, which can then be chosen based on ease of transfection, complement of native G proteins, or other desirable criteria.

A similar result was observed in the case of the serotonin 5-HT_{1D} receptor, which illustrates an extreme case of the variations that can occur in the pharmacological properties of homologous genes in different species. The 5-HT_{1D} receptor is not present in rat cortical membranes; however, the rat brain contains a homologous receptor that exhibits such different pharmacological properties that it has been named a separate serotonin receptor subtype (5-HT_{1B}). Recently, a gene encoding the rat 5-HT_{1B} receptor was isolated and shown to be highly homologous to the human 5-HT_{1D} receptor gene (Adham et al. 1992). This demonstrates that a previous suggestion that the 5-HT_{1B} and 5-HT_{1D} receptors are essentially the same receptor subtype (Hoyer and Middlemiss 1989) was correct. This suggestion was based on the fact that both receptor subtypes show similar distribution in the basal ganglia, similar coupling to adenylate cyclase inhibition, and similar roles as terminal autoreceptors on raphe neurons (Waabber et al. 1990). When a human 5-HT_{1D} receptor clone was expressed in the same mouse fibroblast cell line used to express the 5-HT_{1B} receptor, this human receptor exhibited the binding properties appropriate for the human origin of its gene, rather than the mouse origin of its cell host (Branchek et al. 1991). It appears that no counterexample has been described where the host cell line dictates the pharmacological binding properties of a transfected gene. Protein processing and glycosylation patterns will differ from cell to cell, which may influence receptor kinetics, desensitization phenomena, receptor turnover, and the membrane compartment distribution of the receptor. Nevertheless, it appears that the pharmacological binding properties of 7TM receptors are determined primarily by their gene sequences with a much smaller role (if any) played by their cellular environments.

THE RELATIONSHIP BETWEEN AGONIST AND ANTAGONIST BINDING SITES

G protein-coupled receptors cycle through a complex series of G protein- and ligand-binding affinities. These binding states induce different affinity states for agonist ligands. The availability of transfected cell lines expressing single receptor subtypes now makes it possible to examine these binding states in much greater detail. Two recent studies (Branchek et al. 1990; Teller et al. 1990) have resolved a longstanding controversy in the serotonin receptor field, namely, whether the serotonergic binding site for the agonist 4-bromo-2,5-dimethoxyphenylisopropylamine (DOB) and other related compounds is the high affinity agonist binding state of the 5-HT₂ receptor, which binds antagonist ligands such as ketanserin, or is a separate, closely related receptor subtype. Both interpretations have been advanced (Lyon et al. 1987; Pierce and Peroutka 1989), but the existing data now appear to strongly support the two-site rather than the two-receptor interpretation, as two studies utilizing transfected human (Branchek et al. 1990) and rat (Teller et al. 1990) 5-HT₂ receptor clones have shown. Both studies reached the same conclusion: Transfection of a single cDNA clone into host mammalian cells produced two distinct binding sites (distinct [³H]DOB and [³H]ketanserin binding sites). Addition of guanine nucleotides to these systems reduced the number of agonist high-affinity binding sites with no change, or a slight increase, in the number of antagonist binding sites. Thus, it appears that [³H]DOB and [³H]ketanserin binding sites are distinct ligand-affinity states that exist at different times on the same 5-HT₂ receptor protein. A summary diagram of these affinity states, based on a review article by Freissmuth and colleagues (1989), is provided in figure 1.

The scientific community's increasing molecular understanding of this ligand affinity cycle needs to be better integrated into investigations of receptor function and into drug design programs. Because a complex interaction cycle involving two separate proteins and several forms of guanine nucleotides is involved in agonist binding, the model systems used as templates for drug design must be carefully chosen and carefully adjusted. They must properly mimic the natural processes occurring in those regions of native human brain that are chosen as a target for drug design. Because the types of G proteins, the relative receptor excess (spare receptors), and the amounts of intracellular guanosine diphosphate (GDP) and guanosine triphosphate (GTP) may vary widely in different brain regions, this complexity should be dealt with from the start of a drug design effort. Fortunately, the great freedom of choice of cell hosts and transfection densities that is possible when using cloned human receptors allows just the type of experimental freedom needed to address these issues. It also must be kept in mind that two distinct, but partially overlapping, sets of conformational states are involved in antagonist and agonist binding, and it is important to be sure that the proper mix of the proper states is present in biological screening models.

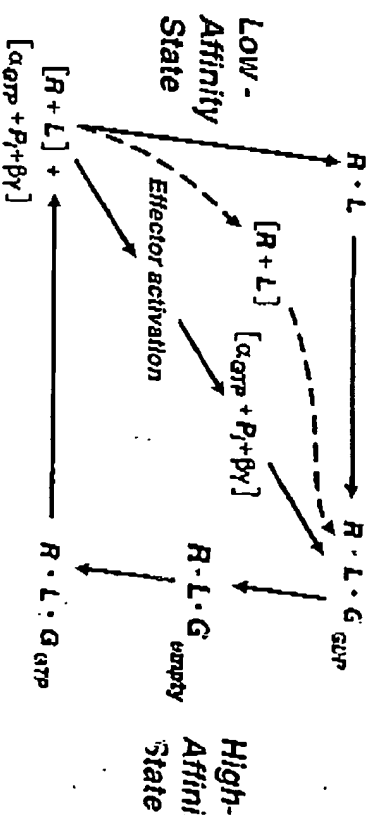


FIGURE 1. Model for the functional activity cycle of G protein-coupled receptors (derived from Freissmuth and colleagues [1989]). Binding of an agonist ligand (L) to the receptor (R) produces a receptor-ligand complex (R·L), which binds a G protein complex. Agonists induce dissociation of guanosine diphosphate (GDP) from this complex, resulting in a "G empty" state, which is the high-affinity agonist-binding state. The G protein can then bind guanosine triphosphate (GTP), which induces dissociation of the α -subunit with GTP bound. This α -subunit can activate effector mechanisms (such as adenylate cyclase or phospholipase C) until such time as the intrinsic GTPase activity of the α -subunit hydrolyses GTP to GDP, starting the cycle again.

AMINO ACID AND PHARMACOLOGICAL HOMOLOGIES BETWEEN RECEPTORS

Receptor researchers have traditionally organized their research groups around a specific neurotransmitter or hormone that activates a set of closely related receptors. For example, serotonin clubs and histamine meetings tend to segregate away from neurokinin researchers, even though all three of these biological mediators activate a group of 7TM receptors with many properties in common. Recent advances in molecular understanding of receptor structures suggest that this approach should change. Receptors of the 7TM superfamily are much more closely related in amino acid sequence and function than had been recognized previously. In addition, Hartig and colleagues (1992) have found that 7TM receptors of the same family (e.g., serotonin receptors) are often less closely related to each other than they are to other 7TM receptors. For example, the serotonin 5-HT_{1A} and 5-HT_{1D} receptors are both more homologous to several adrenergic receptors than they are to either the serotonin 5-HT_{1C} or 5-HT_{2C} receptors (Hartig et al. 1992). Interestingly, this relationship is also

reflected in the pharmacological binding properties of these clones. Many compounds that are classically recognized as adrenergic compounds exhibit moderate-to-high affinity for the 5-HT_{1A} receptor, and a similar relationship holds for the 5-HT_{1D} receptor. For example, yohimbine, which is a classic α_2 antagonist, exhibits a 25-nM affinity for the 5-HT_{1D} receptor (Branchek et al. 1991). This value is only sixfold weaker than its affinity at the α_2A site (Koblika et al. 1987).

These molecular homologies have implications that extend beyond nucleotide sequence comparisons into the field of pharmacology. As is true for all proteins, the form of a neurotransmitter receptor determines its function. As a consequence, the pharmacological binding properties of 7TM receptors often cross family lines in ways that are reflected in the amino acid sequences of these receptors. In general, neurotransmitter receptors that display significant amino acid sequence homologies often display some overlap in the chemical structures that they will bind. Thus, the existence of greater than 50-percent transmembrane amino acid sequence homology should alert researchers to the possibility that receptors from different families may be more closely related in their binding properties than had been previously appreciated. Increased attention to these similarities in sequence and pharmacology could greatly aid the medicinal chemist in efforts to find high-affinity compounds for each receptor subtype and will help establish a much more directed approach to drug design.

REFERENCES

- Adham, N.; Romanienko, P.; Hartig, P.; Weinstanek, R.; and Branchek, T. The rat 5-HT_{1B} receptor is the species homologue of the human 5-HT_{1D} receptor. *Mol Pharmacol* 41:1-7, 1992.
- Branchek, T.; Adham, N.; Macchi, M.; Kao, H.-T.; and Hartig, P.R. [³H]-DOB (4-bromo-2,5-dimethoxyphenylisopropylamine) and [³H]-ketanserin label two affinity states of the cloned human 5-HT₂ receptor. *Mol Pharmacol* 38:681-688, 1990.
- Branchek, T.; Zgombick, J.; Macchi, M.; Hartig, P.; and Weinstanek, R. Cloning and expression of a human 5-HT_{1D} receptor. In: Saxena, P., and Fozard, J.R., eds. *Serotonin: Molecular Biology, Receptors, and Functional Effects*. Basel, Switzerland: Birkhauser, 1991, pp. 21-32.
- Chapot, M.; Eschda, Y.; Marullo, S.; Guillot, J.; Charbit, A.; Strosberg, A.; and Delavier-Klutchko, C. Localization and characterization of three different beta-adrenergic receptors expressed in *Escherichia coli*. *Eur J Biochem* 187:137-144, 1990.
- Freissmuth, M.; Casey, P.; and Gilman, A. G proteins control diverse pathways of transmembrane signaling. *FASEB J* 3:2125-2131, 1989.
- Hartig, P.; Adham, N.; Zgombick, J.; Weinstanek, R.; and Branchek, T. Molecular biology of the 5-HT₁ receptor subfamily. *Drug Dev Res* 26:215-224, 1992.
- Hartig, P.; Kao, H.-T.; Macchi, M.; Adham, N.; Zgombick, J.; Weinstanek, R.; and Branchek, T. The molecular biology of serotonin receptors: An overview. *Neuropsychopharmacology* 3:335-347, 1980.
- Hoyer, D., and Middlemiss, D.N. The pharmacology of the terminal 5-HT autoreceptors in mammalian brain: Evidence for species differences. *Trends Pharmacol Sci* 10:130-132, 1989.
- Hoyer, D.; Pazos, A.; Probst, A.; and Palacios, J.M. Serotonin receptors in the human brain. II. Characterization and autoradiographic localization of 5-HT_{1C} and 5-HT₂ recognition sites. *Brain Res* 376:97-107, 1986.
- Koblika, B.; Maisel, H.; Koblika, T.; Yang-Feng, T.; Francke, U.; Caron, M.; Lefkowitz, R.; and Regan, J. Cloning, sequencing, and expression of the gene coding for the human platelet α_2 -adrenergic receptor. *Science* 238:650-656, 1987.
- Lyons, R.A.; Davis, K.H.; and Teller, M. [³H]-DOB (4-bromo-2,5-dimethoxyphenylisopropylamine) labels guanyl nucleotide-sensitive state of cortical 5-HT₂ receptors. *Mol Pharmacol* 31:194-193, 1987.
- Pazos, A.; Hoyer, D.; and Palacios, J. Mesulergine, a selective serotonin-2 ligand in the rat cortex, does not label these receptors in porcine and human cortex: Evidence for species differences in brain serotonin-2 receptors. *Eur J Pharmacol* 106:531-538, 1984.
- Pierce, P., and Peroutka, S.J. Evidence for distinct 5-hydroxytryptamine₂ receptor binding site subtypes in cortical membrane preparations. *J Neurochem* 52:656-658, 1989.
- Teller, M.; Leonhardt, S.; Weisberg, E.; and Hoffman, B. 4-[¹²⁵I]-do-(2,5-dimethoxy)phenylisopropylamine and [³H]-ketanserin labeling of 5-HT₂ receptors in mammalian cells transfected with a rat 5-HT₂ cDNA: Evidence for multiple states and not multiple 5-HT₂ receptor subtypes. *Mol Pharmacol* 38:594-598, 1990.
- Waerber, C.; Schoeffter, P.; Hoyer, D.; and Palacios, J.M. The serotonin 5-HT_{1D} receptor: A progress review. *Neurochem Res* 15:567-582, 1990.

AUTHOR

Paul R. Hartig, Ph.D.
Vice President for Research
Synaptic Pharmaceuticals (formerly Neurogenetic Corp.)
215 College Road
Paramus, NY 07652

A SUPPLEMENT TO

The Journal of Nuclear Medicine

JNM

Volume 36, Number 6 • June 1995

Molecular Nuclear Medicine



Backbone structure of the human growth hormone-receptor complex.
See page 15S.



The Official Publication of
The Society of Nuclear Medicine, Inc.

BEST AVAILABLE COPY

Volume 36, Number 6 • June 1995 • Supplement

JNM



Official Publication of
The Society of Nuclear Medicine

Molecular Nuclear Medicine

- 1S Introduction**
Richard C. Reba
- 2S Molecular Nuclear Medicine: From Genotype to Phenotype via Chemotype**
Henry N. Wagner, Jr.
- 5S Designing a Molecular Probe for Muscarinic Acetylcholine Receptor (mAChR) Imaging**
William C. Eckelman
- 8S Designing Steroid Receptor-Based Radiotracers to Image Breast and Prostate Tumors**
John A. Katzenellenbogen
- 14S Structure of the Growth Hormone-Receptor Complex and Mechanism of Receptor Signaling**
Anthony A. Kossiakoff
- 17S Structure and Functional Analysis of G Protein-Coupled Receptors and Potential Diagnostic Ligands**
Claire M. Fraser
- 22S Recurring Genetic Aberrations in Cancer Cells: Chromosomes as Potential Targets for Nuclear Medicine Imaging**
Janet D. Rowley
- 25S Structure of the BPV-1 E2 DNA-Binding Domain Bound to Its DNA Target**
Rashmi S. Hegde
- 28S Structure-Aided Drug Design: Crystallography and Computational Approaches**
Dagmar Ringe



Publication of this supplement was supported by grant DE-FG02-93ER61671 from the Office of Health and Environmental Research, U.S. Department of Energy. This grant was administered by the American College of Nuclear Physicians, Washington, D.C. These materials were edited and coordinated by Linda E. Ketchum, Ketchum InfoMedia, Inc., New York, NY.

The opinions expressed in this publication are those of the authors and are not attributable to the publisher, editor-in-chief or editorial board of *The Journal of Nuclear Medicine*.

The image on the cover is reprinted with permission from: de Vos AM, Kossiakoff AA. Receptor action and interaction. *Current Opinions in Structural Biology* 1992;2:852-858.

Int

Guest

Depart

J Nucl

Adv
the spe
for a p
takes y
of mol
publis
Supp
Depart
Resear
promis
and cl
design
In th
opens

Rece
For c
Rd., Ger

Intro

Structure and Functional Analysis of G Protein-Coupled Receptors and Potential Diagnostic Ligands

Claire M. Fraser

The Institute for Genomic Research, Gaithersburg, Maryland

G protein-coupled receptors are a diverse class of proteins that mediate signal transduction across the plasma membrane. More than 200 receptors in this extended gene family have been cloned, and comparison of the deduced amino-acid sequences indicates that these proteins have marked homology and share a common membrane topology consisting of seven transmembrane helices. Although there is considerable variability in the physiologic ligands responsible for receptor activation, all receptors in this group interact with trimeric, guanine nucleotide-binding proteins to initiate signaling cascades in the cell cytosol. To investigate the structural motifs responsible for ligand binding, we have established a model system to express heterologously human G protein-coupled receptors in a mammalian cell line. This experimental system allows each receptor subtype to be studied in isolation and provides a direct means to link receptor activation to a particular second messenger cascade. Furthermore, the efficacy and specificity of new pharmaceuticals can now be evaluated readily with cloned human receptors, eliminating the need for animal tissues. We have used this expression system in conjunction with an experimental strategy of site-directed mutagenesis to identify amino-acid residues that have a functional role in ligand binding. Because of the strong homology that exists within this family of receptor proteins, the results of this work are applicable to other systems and, therefore, can help to establish a more complete understanding of ligand-receptor interactions. This combined molecular and biochemical approach to the study of G protein-coupled receptors can pave the way for the development of isoform-specific ligands that may be used for radionuclide imaging and therapy.

J Nucl Med 1995; 36(Suppl):17S-21S

Cell surface receptors are integral membrane proteins that connect external stimuli to biochemical changes within the cell. These proteins can be grouped into three

superfamilies based on their primary structures and mechanisms of action:

1. Receptors that bind growth factors.
2. Ligand-gated ion channels, such as the nicotinic, gamma-aminobutyric acid (GABA) and glycine receptors.
3. Receptors that interface with guanine nucleotide-binding regulatory proteins.

The third group, G protein-coupled receptors, is a diverse collection of proteins that includes distinct receptor subfamilies activated by peptide hormones, neurotransmitters, or environmental stimuli (Table 1).

Although G protein-coupled receptors have different physiologic activators, they have two unifying characteristics:

1. Each protein contains seven stretches of high hydrophobicity that appear to form membrane-spanning segments. Therefore, all receptors in this class are thought to share a similar membrane topology, analogous to the structure of bacteriorhodopsin (Fig. 1). This proposed topology has been confirmed for both rhodopsin (1) and the beta-adrenergic receptor (2) through the use of antipeptide antibodies directed against specific regions of the receptor protein.
2. In each system, receptor stimulation causes the activation of a trimeric G protein on the cytosolic surface of the plasmalemma (3). Interaction with a G protein, therefore, is the common primary step of each signalling cascade. In the activated state, the G alpha subunit dissociates from the beta-gamma complex. Diversification of the biochemical response is caused by the subsequent modulation of additional effector enzymes by G alpha (Fig. 2). These downstream elements may include: phospholipases A, C, or D; adenylate or guanylate cyclase; or other proteins, such as ion channels.

Pharmacologic analysis over the past 10 to 15 yr suggested that many receptor classes were, in fact, a group of closely related isoforms. This premise was supported by the observation that a specific ligand, such as acetylcholine, could elicit distinct biochemical responses in different tissues. Moreover, the sensitivity of receptors to

Received Jan. 11, 1995; accepted Feb. 27, 1995.

For correspondence or reprints contact: Claire M. Fraser, PhD, The Institute for Genomic Research, 932 Clopper Road, Gaithersburg, MD 20878.

Financial support for this work was provided by The National Institute on Alcohol Abuse and Alcoholism.

TABLE 1
Membrane Receptors That Interact with G Proteins

Peptide Hormone Receptors	Glycoprotein Hormone Receptors
Angiotensin	Choriogonadotropin
Adrenocorticotropin (ACTH)	Follicle-stimulating hormone (FSH)
Antidiuretic hormone	Thyrotropin (TSH)
Bombesin	
Bradykinin	Neurotransmitter Receptors
Calcitonin	Adenosine
Cholecystokinin (CCK)	Adenosine triphosphate (ATP)
C5a anaphylatoxin	Alpha-Adrenergic
Corticotropin-releasing hormone (CRF)	Beta-Adrenergic
Endothelin	Dopamine
Gastrin	Gamma-aminobutyric acid (GABA)
Glucagon	Glutamate
Glucagon-like peptide	Histamine
Gonadotropin-releasing hormone (GnRH)	Muscarinic acetylcholine
Growth hormone-releasing hormone (GRF)	Octopamine
Interleukin-8	Serotonin
Kinins (bradykinin, substances P and K)	Tyramine
Leutinizing hormone (LH)	
Melanocortin	Sensory Systems
Melanocyte-stimulating hormone (MSH)	Vision (rhodopsins)
N-formyl peptide	Olfaction
Neuropeptide tyrosine (NPY)	
Neurotensin	Other Agents
Opiates	Cannabinoids
Oxytocin	Immunoglobulin E (IgE)
Parathyroid hormone	Mas oncogene
Pituitary adenylate cyclase-activating protein	Platelet-activating factor
Secretin	Prostanoids
Somatostatin	Thrombin
Thyrotropin-releasing hormone (TRH)	
Vasoactive intestinal polypeptide (VIP)	
Vasopressin	

agonists or antagonists varied with the experimental material. These early observations have been confirmed with the cloning of over 200 genes that encode G protein-coupled receptors (4). Comparison of the predicted protein sequences illustrated that most receptors are part of a multigene family that may include as many as six isoforms (4). In addition, low-stringency screening and the application of new molecular cloning techniques have led to the identification of novel receptor subtypes that were not previously anticipated from pharmacologic studies.

These findings highlight one of the most challenging problems in the development of useful drugs for radionuclide imaging and therapy: How can pharmaceuticals be designed and tested that are specific for a particular receptor isoform? To address this problem adequately, it is essential to answer the following questions: Which second messenger cascade is elicited by a particular receptor subtype? How is the response affected by different agonists? It is equally apparent, from the heterogeneity of receptor

proteins in vivo, that the answers to these questions will require the development of new experimental systems that can ascertain the properties of each receptor subtype.

HETEROLOGOUS EXPRESSION

We have used an experimental system in which cloned G protein-coupled receptors are stably transfected and expressed in a mammalian cell line (5). Heterologous expression of receptor proteins has two major advantages: analysis of a single receptor subtype in isolation and study of drug interactions with human receptors, eliminating the need for animal tissues in drug screening protocols. Although our research has focused on the muscarinic acetylcholine receptor, the observations concerning receptor-ligand interaction are relevant to any one of a number of G protein-coupled receptors. Hence, within this gene superfamily, there exists some commonality of structure and function.

Five distinct muscarinic receptor genes have been cloned and sequenced (6) and have been designated m1 through m5. The m1, m3 and m5 subtypes preferentially stimulate phosphoinositide hydrolysis in response

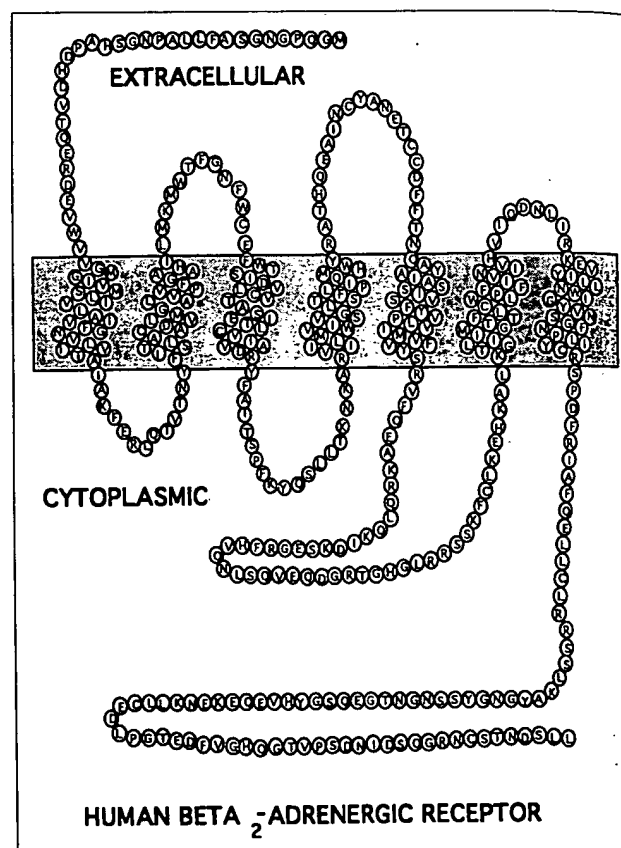


FIGURE 1. Schematic illustration of cell membrane topology of G protein-coupled receptors with seven stretches of membrane-spanning segments with high hydrophobicity. [Reprinted with permission from: Lee NH, Fraser CM. Identifying the functional domains of G protein-coupled-receptors. In: Krogsaard-Larsen P, Christensen S, Kofod H, eds. *News leads and targets in drug research: Alfred Benzon symposium no. 33*. Copenhagen: Munksgaard; 1992:187-199.]

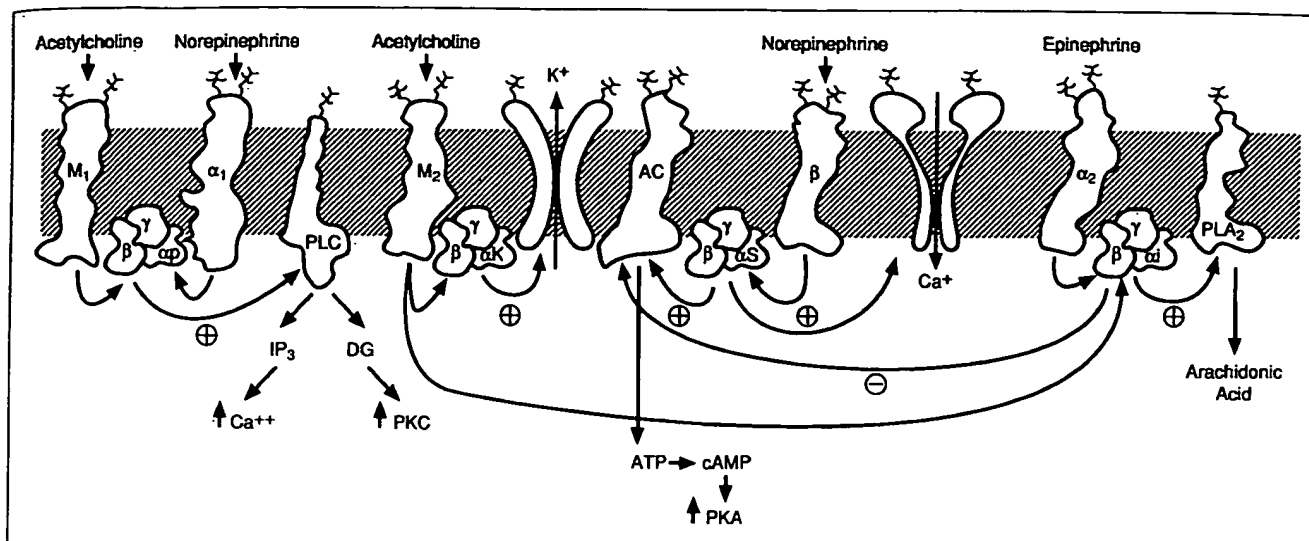


FIGURE 2. Schematic illustration of the signal transduction mechanisms common to G protein-coupled receptors. Agonist binding to G protein-coupled receptors promotes receptor coupling to various heterotrimeric G proteins, which catalyze the exchange of bound guanosine diphosphate (GDP) for guanosine triphosphate (GTP) on the G protein alpha subunits. Binding of GTP to the alpha subunits results in dissociation of the G protein heterotrimer complex. Depending on the receptor and the G protein with which it is associated, the $G\alpha$ -GTP subunit activates (+) or inhibits (-) one or more intracellular effector enzymes, leading to metabolic changes in the cell. Acetylcholine binds to subtypes of muscarinic acetylcholine receptors indicated as M1 and M2. Norepinephrine and epinephrine bind to subtypes of alpha- (α) and beta- (β) adrenergic receptors. The heterotrimeric G proteins are composed of alpha, beta and gamma subunits. Effector enzymes stimulated or inhibited by G protein-coupled receptors include: (a) adenylate cyclase (AC), which converts adenosine triphosphate (ATP) to cyclic adenosine monophosphate (cAMP) and activates protein kinase A (PKA); (b) phospholipase C (PLC), which hydrolyzes inositol phospholipids to produce inositol phosphates (IP3) and diacylglycerol (DG) [inositol phosphates increase the levels of intracellular calcium and diacylglycerol stimulates protein kinase C (PKC) activity]; (c) phospholipase A2, which hydrolyzes membrane lipids to produce arachidonic acid; and (d) various ion channels, which modulate ion flow across the cell membrane.

to agonist binding, whereas the m2 and m4 subtypes preferentially inhibit adenylate cyclase (6). Each receptor however, can activate more than one intracellular signaling pathway under the appropriate conditions. For example, the phosphoinositide-coupled muscarinic receptors have been shown to mediate an increase in intracellular cyclic adenosine monophosphate (cAMP) and also may stimulate the release of arachidonic acid from membranes (7,8).

Equally interesting are the differences observed in the magnitude of the responses elicited by receptors that activate the same second messenger cascade (7). The m1 and m3 isoforms both stimulate the phosphoinositide pathway. Yet, comparison of the m1 and m3 subtypes, expressed in chinese hamster ovary (CHO) cells, illustrated that the phosphoinositide response evoked by agonist binding to the m1 muscarinic receptor was always greater than that observed with the m3 receptor. These differences were not due to dissimilarities in the level of gene expression since both receptors were present at the plasma membrane in equivalent densities. It is not clear whether this difference reflects the coupling of these two receptor subtypes to distinct G proteins or a differential coupling to a single G protein. Nevertheless, these observations suggest that there may be physiologically relevant differences in the coupling of receptor isoforms to the same biochemical pathway.

We have also observed agonist-specific activation of intracellular signalling pathways (9). Three muscarinic agonists—carbachol, pilocarpine, and AF102B—were examined for their ability to stimulate phosphoinositide hydrolysis, cAMP production and arachidonic acid release from CHO cells transfected with the m1 muscarinic receptor. Carbachol and pilocarpine produced maximal stimulation of phosphoinositide hydrolysis. This response was greater than the phosphoinositide hydrolysis elicited by AF102B. Similar results were found when arachidonic acid release was monitored. In contrast, only carbachol produced an increase in the level of cytosolic cAMP, whereas pilocarpine and AF102B had no effect on this pathway. These data support findings from other studies with m1 muscarinic receptors (10).

Comparison of the chemical structure of these agonists suggests one plausible explanation for the diverse response of the m1 receptor: Carbachol is the compound with the most flexibility since it can assume four or five conformational states that have a similar energy minima (9). Multiple conformational states may allow distinct ligand-receptor interaction, which might account for the diversity observed in the biochemical response. Pilocarpine and AF102B, on the other hand, have more rigid chemical structures, which may limit the ability of these compounds to stimulate completely the m1 receptor. Interestingly, it has been postulated

that the ability of AF102B to function as a partial agonist may be important therapeutically because patients may not develop tolerance to this compound (9). This drug is currently in clinical trials in Japan as a treatment for Alzheimer's disease.

DOWNREGULATION

The phenomenon of receptor downregulation and its relation to drug tolerance is a serious problem in the development of therapeutics. It has been shown that long-term incubation of many G protein-coupled receptors with an agonist produces a reduction in the number of receptor proteins at the cell surface (4). This process is called receptor downregulation. One clinical manifestation of this phenomenon is tachyphylaxis, observed in asthma patients who use beta-adrenergic agonists, such as bronchial dilators. Chronic administration of these beta-adrenergic agonists may cause patients eventually to become refractory to the agonist's beneficial effects.

We have examined the biochemical and molecular features of receptor downregulation in CHO cells transfected with the m1 muscarinic acetylcholine receptor, the α_2 -adrenergic receptor, and the β_2 -adrenergic receptor (11). Following 24-hr incubation with carbachol, a muscarinic agonist, transfected CHO cells showed a reduction in the magnitude of phosphoinositide hydrolysis elicited by reapplication of carbachol in comparison to cells that had no prior carbachol exposure. Interestingly, the addition of isoproterenol, a beta-adrenergic agonist, also caused a reduction in the carbachol-induced phosphoinositide response of the muscarinic receptor. Therefore, long-term activation of either G protein-coupled receptor (the muscarinic or the beta-adrenergic receptor) caused downregulation of the muscarinic receptor. The reduction in receptor density at the cell surface correlated with a decrease in the level of messenger ribonucleic acid (mRNA) specific for the muscarinic receptor.

These observations indicate that receptor downregulation is in part a biochemical feedback process that reduces the level of gene transcription in response to receptor stimulation. These findings may be important in the long-term therapy of diseases with some of these agonists and represent a possible utility for radionuclide imaging as a technique to monitor changes in receptor levels in target tissues.

SITE-DIRECTED MUTAGENESIS

Along with the biochemical analysis of G protein-coupled receptors, we have used this heterologous expression system to identify structures within receptor protein that have functional importance (4). These studies employed an experimental strategy of site-directed mutagenesis followed by expression of the mutant receptor protein in transfected cells to define regions responsible for ligand binding and receptor activation by agonists. A similar

strategy has been utilized in other laboratories to determine receptor domains that interact with G proteins (4) and amino acid residues that undergo post-translational modifications, such as glycosylation, which may be essential for normal receptor function (4).

We have focused on amino acid residues that are highly conserved among all G protein-coupled receptors and positioned toward the extracellular membrane surface where ligand-binding is thought to occur. One caveat to this experimental approach is the possibility that a point mutation will cause a large-scale conformational change in the protein. In such a case, receptor inactivity may be caused by protein misfolding and because the mutated residue had a critical role in receptor function. To minimize this problem, we have made conservative amino acid substitutions, replacing the original residue with one of similar size and/or hydrophobicity.

One of the striking features of most receptors in this family is the presence of two conserved cysteine residues (4), one in the extracellular loop between helices II and III and a second in the extracellular loop between helices IV and V (Fig. 1). Biochemical evidence from a number of G protein-coupled receptor systems has suggested that these cysteines may form a disulfide linkage, covalently connecting the two extracellular loops (4,12). We have made mutations at each position, changing the cysteine to a serine residue, in the muscarinic acetylcholine receptor. In each case, the transfected cells expressed the mutant receptor, as evidenced by Northern analysis (13), but no agonist-mediated increase in phosphoinositide hydrolysis could be observed.

These results confirm the earlier biochemical studies and also suggest that disulfide formation is essential for maintaining the correct protein conformation required for recognizing ligands and receptor activation.

The precise location of the ligand binding-site has yet to be determined. Earlier work implied that ligands were bound within the transmembrane domains, since large deletions in the beta-adrenergic receptor could be made in either the extracellular or cytosolic loops without affecting ligand-receptor association (14). In light of these findings, we began to look at these domains and specific amino acids within the transmembrane helices, asking whether these residues had a role in ligand binding. Alignment of the deduced amino acid sequences from a number of G protein-coupled receptors revealed that a single aspartic acid residue within helix III is absolutely conserved among all receptors that bind ligands with a positively charged nitrogen. Examples include the following proteins: muscarinic receptors that bind acetylcholine, adrenergic receptors that bind epinephrine and norepinephrine, dopamine receptors, serotonin receptors, and histamine receptors. Moreover, it has been postulated that this negatively charged aspartic acid may play a role in binding the positively charged nitrogen common among these ligands (15).

We have examined the role of this aspartic acid by mutating it to an asparagine in beta- and alpha-adrenergic receptors and in the muscarinic acetylcholine receptor. All three mutant receptors were unable to bind radiolabeled ligands, whereas the wildtype proteins displayed a high-affinity, saturable binding of the appropriate compound (16). Our findings corroborate results published by Hulme et al. (17), which determined that this same aspartic acid residue in the muscarinic receptor was covalently linked to the radioactive affinity-probe, propylbenzilylcholine mustard.

Work from our laboratory and others have also implicated transmembrane threonine, tyrosine and cysteine residues in agonist binding, although it is not yet known whether any of these residues directly participate in receptor-ligand interactions (18,19). All of these amino acids are located in the same plane of the membrane, within one to two turns of the alpha helix from the extracellular surface, supporting the idea that agonist binding may occur within the upper third of the transmembrane helices.

CONCLUSION

A combined approach of heterologous gene expression and site-directed mutagenesis provides a starting point for future structure-function analysis of G protein-coupled receptors. These studies, along with efforts toward obtaining a receptor crystal structure, may make it possible to design more selective ligands for radionuclide imaging and therapy.

REFERENCES

1. Applebury ML, Hargrave PA. Molecular biology of the visual pigments. *Vision Res* 1986;26:1881-1895.
2. Wang HY, Lipfert L, Malbon CC, Bahouth S. Site-directed anti-peptide antibodies define the topography of the β 2-adrenergic receptor. *J Biol Chem* 1989;264:14424-14431.
3. Lyengar R, Birnbaumer L, eds. *G proteins*. New York: Academic Press; 1990.
4. Savarese TM, Fraser CM. In vitro mutagenesis and the search for structure-function relationships among G protein-coupled receptors. *Biochem J* 1992;283:1-19.
5. Fraser CM. Expression of receptor genes in cultured cells. In: Hulme EC ed. *Receptors: a practical approach*. Oxford, UK: IRL Press; 1990:263-275.
6. Hulme EC, Birdsall NJM, Buckley NJ. Muscarinic receptor subtypes. *Ann Rev Pharmacol Toxicol* 1990;30:633-673.
7. Buck MA, Fraser CM. Muscarinic acetylcholine receptor subtypes which selectively couple to phospholipase C: Pharmacological and biochemical properties. *Biochem Biophys Res Commun* 1990;173:666-672.
8. Conklin BR, Brann MR, Buckley NJ, Ma AL, Bonner TI, Axelrod J. Stimulation of arachidonic acid release and inhibition of mitogenesis by cloned genes for muscarinic receptor subtypes stably expressed in A9 L cells. *Proc Natl Acad Sci USA* 1988;85:8698-8702.
9. Gurwitz D, Haring R, Heldman E, Fraser CM, Manor D, Fisher A. Discrete activation of transduction pathways associated with acetylcholine m1 receptor by several muscarinic ligands. *Eur J Pharmacol* 1994;267:21-31.
10. Felder CC, Kantertnan RY, Ma AL, Axelrod J. A transfected m1 muscarinic acetylcholine receptor stimulates adenylate cyclase via phosphatidylinositol hydrolysis. *J Biol Chem* 1989;264:20356-20362.
11. Lee NH, Fraser CM. Cross-talk between m1 muscarinic acetylcholine and β 2-adrenergic receptors. *J Biol Chem* 1993;268:7949-7957.
12. Curtis CAM, Wheatley M, Bansal S, et al. Propylbenzilylcholine mustard labels an acidic residue in transmembrane helix 3 of the muscarinic receptor. *J Biol Chem* 1989;264:489-495.
13. Savarese TM, Wang C-D, Fraser CM. Analysis of the function of conserved cysteine residues in m1 muscarinic acetylcholine receptor function using site-directed mutagenesis and permanent expression in CHO cells. *J Biol Chem* 1992;267:11439-11448.
14. Dixon RAF, Sigal IS, Rands E, et al. Ligand binding to the β -adrenergic receptor involves its rhodopsin-like core. *Nature* 1987;326:73-77.
15. Strader CD, Sigal IS, Register RB, Candelore MR, Rands E, Dixon RAF. Identification of residues required for ligand binding to the β -adrenergic receptor. *Proc Natl Acad Sci USA* 1987;84:4384-4388.
16. Fraser CM, Wang C-D, Robinson DA, Gocayne JD, Venter JC. Site-directed mutagenesis of m1 muscarinic receptors: conserved aspartic acids play important roles in receptor function. *Mol Pharmacol* 1989;36:840-847.
17. Kurtenbach E, Curtis CAM, Pedder EK, Aitken A, Harris ACM, Hulme EC. Muscarinic acetylcholine receptors. Peptide sequencing identifies residues involved in antagonist binding and disulfide bond formation. *J Biol Chem* 1990;265:13702-13708.
18. Wess J, Gdula D, Brann MR. Site-directed mutagenesis of the m3 muscarinic receptor: Identification of a series of threonine and tyrosine residues involved in agonist but not antagonist binding. *EMBO J* 1991;10:3729-3734.
19. Muzzin P, Revelli J-P, Kuhne F, Gocayne JD, McCombie WR, Venter JC. A novel adipocyte β -adrenergic receptor: molecular cloning and specific down-regulation in obesity. *J Biol Chem* 1991;266:24053-24058.
20. Lee NH, Fraser CM. Identifying the functional domains of G protein-coupled-receptors. In: Krogsgaard-Larsen P, Christensen S, Kofod H, eds. *News leads and targets in drug research: Alfred Benzon symposium no. 33*. Copenhagen: Munksgaard; 1992:187-199.

A Regulatory Cascade of the Nuclear Receptors FXR, SHP-1, and LRH-1 Represses Bile Acid Biosynthesis

Bryan Goodwin,* Stacey A. Jones,* Roger R. Price,* Michael A. Watson,* David D. McKee,* Linda B. Moore,* Cristin Galardi,* Joan G. Wilson,† Michael C. Lewis,† Matthew E. Roth,§ Patrick R. Maloney,‡ Timothy M. Willson,‡ and Steven A. Kliewer*[¶]

*Department of Molecular Endocrinology

†Department of Metabolic Diseases

‡Department of Medicinal Chemistry

Glaxo Wellcome Research and Development
Research Triangle Park, North Carolina 27709

§CuraGen Corporation

New Haven, Connecticut 06511

Summary

Bile acids repress the transcription of cytochrome P450 7A1 (CYP7A1), which catalyzes the rate-limiting step in bile acid biosynthesis. Although bile acids activate the farnesoid X receptor (FXR), the mechanism underlying bile acid-mediated repression of CYP7A1 remained unclear. We have used a potent, nonsteroidal FXR ligand to show that FXR induces expression of small heterodimer partner 1 (SHP-1), an atypical member of the nuclear receptor family that lacks a DNA-binding domain. SHP-1 represses expression of CYP7A1 by inhibiting the activity of liver receptor homolog 1 (LRH-1), an orphan nuclear receptor that is known to regulate CYP7A1 expression positively. This bile acid-activated regulatory cascade provides a molecular basis for the coordinate suppression of CYP7A1 and other genes involved in bile acid biosynthesis.

Introduction

Cholesterol is essential for a number of cellular functions, including membrane biogenesis and steroid hormone and bile acid biosynthesis. However, in excess, cholesterol can contribute to disease processes such as atherosclerosis and gallstone formation. Therefore, cholesterol biosynthesis and catabolism must be coordinately regulated. The metabolism of cholesterol to bile acids represents a major pathway for its elimination from the body, accounting for approximately half of daily excretion. Cytochrome P450 7A (CYP7A1) is a liver-specific enzyme that catalyzes the first and rate-limiting step in one of the two pathways for bile acid biosynthesis (Chiang, 1998; Russell and Setchell, 1992). The gene encoding CYP7A1 is regulated by a variety of small, lipophilic molecules, including steroid and thyroid hormones, cholesterol, and bile acids. Notably, CYP7A1 expression is stimulated by cholesterol feeding and repressed by bile acids. Thus, CYP7A1 is under both feedforward and feedback regulation.

CYP7A1 expression is regulated by several members

of the nuclear receptor superfamily of ligand-activated transcription factors (Chiang, 1998; Gustafsson, 1999; Russell, 1999). Recently, two nuclear receptors, the liver X receptor α (LXR α ; NR1H3) (Apfel et al., 1994; Willy et al., 1995) and farnesoid X receptor (FXR; NR1H4) (Forman et al., 1995; Seol et al., 1995), were implicated in the feedforward and feedback regulation of CYP7A1, respectively (Peet et al., 1998; Russell, 1999). Both LXR α and FXR are abundantly expressed in the liver and bind to their cognate hormone response elements as heterodimers with the 9-*cis* retinoic acid receptor RXR (Mangelsdorf and Evans, 1995). LXR α is activated by the cholesterol derivative 24,25(S)-epoxycholesterol and binds to a response element in the CYP7A1 promoter (Lehmann et al., 1997). Mice lacking LXR α do not induce CYP7A1 expression in response to cholesterol feeding (Peet et al., 1998). Moreover, these animals accumulate massive amounts of cholesterol in their livers when fed a high cholesterol diet. These data establish LXR α as the cholesterol sensor responsible for feedforward regulation of CYP7A1 expression.

Bile acids stimulate the expression of genes involved in bile acid transport, such as the intestinal bile acid-binding protein (*I-BABP*), and repress CYP7A1 and other genes encoding enzymes involved in bile acid biosynthesis, such as CYP8B1, which converts chenodeoxycholic acid (CDCA) to cholic acid, and CYP27, which catalyzes the first step in the alternative, "acidic" pathway for bile acid synthesis (Russell and Setchell, 1992; Javitt, 1994; Russell, 1999). Recently, FXR was shown to be a bile acid receptor (Wang et al., 1996; Makishima et al., 1999; Parks et al., 1999). Several different bile acids, including CDCA and its glycine and taurine conjugates, bind and activate FXR at physiologic concentrations. Moreover, FXR response elements (FXREs) were identified in both the mouse and human *I-BABP* promoters (Grober et al., 1999; Makishima et al., 1999), which provided strong evidence that FXR mediates the positive effects of bile acids on *I-BABP* expression. Notably, the rank order of bile acids that activate FXR correlates with that for repression of CYP7A1 in a hepatocyte-derived cell line (Makishima et al., 1999). These data suggested that FXR also has a role in the negative effects of bile acids on gene expression. However, since the region of the CYP7A1 promoter that is necessary for bile acid-mediated repression lacks a strong FXR-binding site (Chiang and Stroup, 1994; Chiang et al., 2000), it seemed unlikely that this repression was a direct effect of FXR. Thus, the molecular mechanism for bile acid-mediated repression of CYP7A1 remained in question.

In this report, we have used a potent, nonsteroidal FXR ligand to demonstrate that FXR regulates the hepatic expression of small heterodimer partner 1 (SHP-1; NR0B2), an atypical, orphan member of the nuclear receptor family that lacks a DNA-binding domain (Seol et al., 1996). SHP-1 has been shown to bind to other nuclear receptors and to repress their transcriptional activities (Seol et al., 1996; Masuda et al., 1997; Johansson et al., 1999; Lee et al., 2000). We show that SHP-1 represses the CYP7A1 promoter through interaction with liver receptor homolog 1 (LRH-1; NR5A2), an orphan nuclear receptor that binds as a monomer to a response

[¶]To whom correspondence should be addressed (e-mail: sak15922@glaxowellcome.com).

element in the *CYP7A1* promoter and activates transcription (Becker-Andre et al., 1993; Galameau et al., 1996; Nitta et al., 1999). LRH-1 is a mammalian homolog of the *Drosophila* fushi tarazu F1 gene product, which regulates *Drosophila* metamorphosis (Lavorgna et al., 1991; Broadus et al., 1999). Our findings define a novel regulatory cascade of three orphan nuclear receptors that provides a molecular basis for the coordinate repression of gene expression by bile acids.

Results

Identification of GW4064 as a Potent, Selective FXR Activator

FXR was recently shown to be a receptor for CDCA as well as other bile acids (Makishima et al., 1999; Parks et al., 1999; Wang et al., 1999). However, these compounds bind to FXR with only micromolar affinities and at these concentrations also interact with other proteins, including bile acid-binding proteins and transporters. We sought to identify a potent, selective FXR ligand for use as a chemical tool in elucidating the genes regulated by FXR. Combinatorial libraries of compounds were screened using a ligand-sensing fluorescence resonance energy transfer assay that detects interactions between FXR and a peptide derived from the steroid receptor coactivator 1 (SRC-1) as previously described (Parks et al., 1999). Among the compounds that promoted an interaction between FXR and SRC-1 was the isoxazole GW4064 (Figure 1A), which bound to FXR with a half-maximal effective concentration (EC_{50}) of 15 nM (Maloney et al., 2000). GW4064 activated mouse and human FXR with EC_{50} values of 80 and 90 nM, respectively, in CV-1 cells transfected with FXR expression vectors and a reporter plasmid containing two copies of an established FXR response element (FXRE) derived from the *Drosophila* heat shock protein 27 (hsp27) promoter (Forman et al., 1995) (Figure 1B). Thus, GW4064 is ~1000-fold more potent than CDCA in activating FXR in CV-1 cells (Figure 1B).

GW4064 was tested for selectivity against a panel of nuclear receptors. CV-1 cells were transfected with expression plasmids for various nuclear receptor-GAL4 chimeras and the reporter plasmid (*UAS*)₂-tk-CAT as previously described (Parks et al., 1999). GW4064 activated only the FXR-GAL4 chimera (Figure 1C). Thus, GW4064 is a highly selective activator of FXR.

FXR Regulates *SHP-1* Expression in the Liver

GW4064 was exploited as a chemical tool to identify the genes regulated by FXR in the liver. Male Fisher rats were treated for 7 days with GW4064 or vehicle alone (methyl cellulose). Following treatment, RNA was prepared from the livers of GW4064- and vehicle-treated animals, and genes that were either induced or repressed by GW4064 treatment were determined using CuraGen GeneCalling™ differential gene expression technology (Shimkets et al., 1999). A comprehensive list of the liver genes regulated by GW4064 will be published elsewhere. Interestingly, the gene that was most strongly induced by GW4064 treatment was that encoding the orphan nuclear receptor *SHP-1*. Northern analysis showed that *SHP-1* expression was increased ~6-fold in the livers of GW4064-treated rats relative to vehicle-treated animals (Figure 2A).

Bile acids are known to repress the expression of

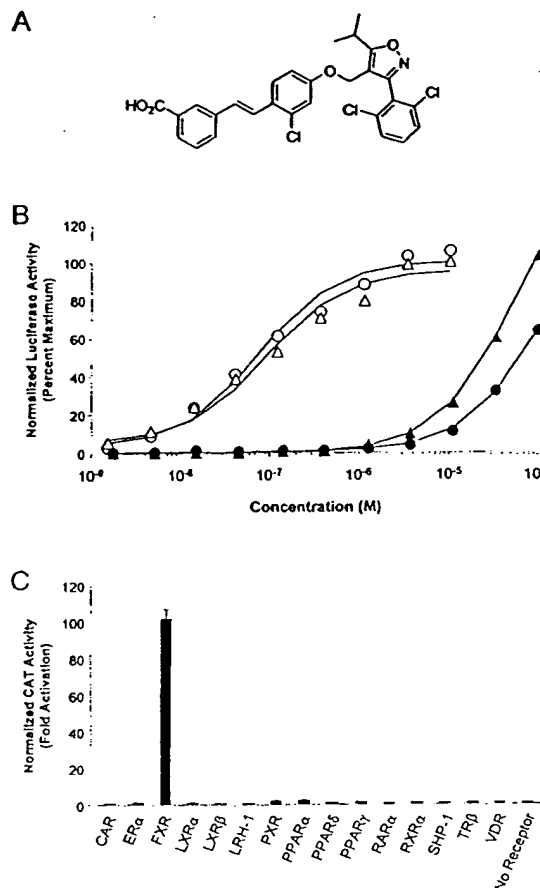
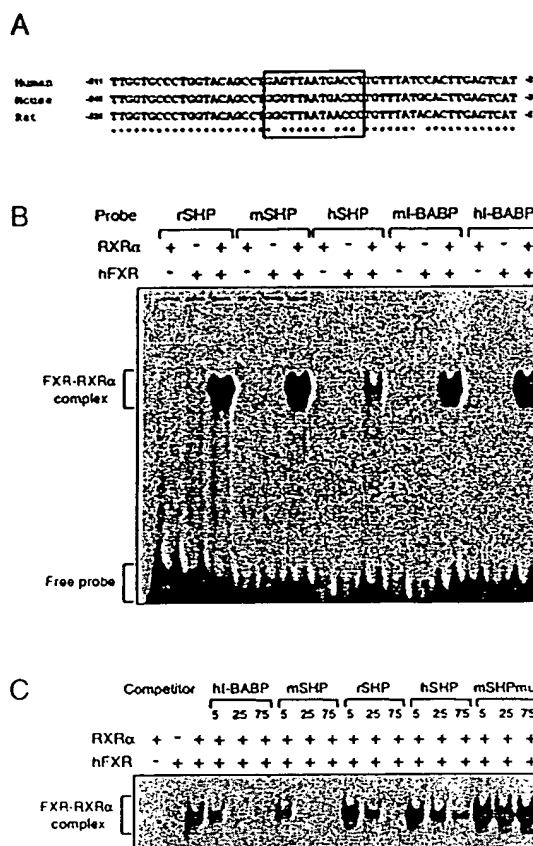


Figure 1. GW4064 Is a Potent, Selective Activator of FXR

(A) Chemical structure of GW4064. (B) CV-1 cells were transfected with expression plasmids for human or mouse FXR and the (hsp70EcRE)₂-tk-LUC reporter plasmid containing two copies of the hsp70 ecdysone response element upstream of the thymidine kinase (tk) promoter and luciferase gene. Transfected cells were treated with the indicated concentrations of either GW4064 or CDCA. Open circles, mouse FXR and GW4064; open triangles, human FXR and GW4064; closed circles, mouse FXR and CDCA; closed triangles, human FXR and CDCA. Data points represent the mean of assays performed in triplicate. (C) CV-1 cells were transfected with expression vectors for various GAL4-nuclear receptor ligand-binding domain chimeras and the reporter plasmid (*UAS*)₂-tk-CAT. Transfected cells were treated with 1 μ M GW4064. Data represent the mean of assays performed in triplicate \pm S.D.

CYP7A1 as part of a regulatory feedback loop that controls the rate of their biosynthesis from cholesterol (Russell and Setchell, 1992; Russell, 1999). Two recent studies implicate FXR in the repression of *CYP7A1* (Makishima et al., 1999; Wang et al., 1999), although the molecular mechanisms have remained unclear since the *CYP7A1* promoter does not contain a consensus FXRE (Chiang et al., 2000). In parallel with our analysis of *SHP-1* expression, we examined whether GW4064 treatment resulted in decreased *CYP7A1* expression in male Fisher rats. Rats treated with GW4064 showed a substantial decrease in *CYP7A1* mRNA levels (~4-fold, Figure 2A). Thus, GW4064 mimics the well documented



(C) Electrophoretic mobility-shift assays were performed with *in vitro* synthesized human FXR and/or human RXR α , a [32 P]-labeled oligonucleotide containing the human *I-BABP* FXRE, and either a 5-, 25-, or 75-fold excess of unlabeled oligonucleotides containing the IR1 motif from the human *I-BABP* promoter, the mouse, rat, or human *SHP-1* promoters, or a mutated derivative of the mouse *SHP-1* IR1 motif (mSHPmut). The position of the shifted FXR/RXR α complex is indicated.

(C) Total RNA was prepared from primary human hepatocytes treated for 48 hr with the indicated concentrations of CDCA. Northern analysis was performed using probes for human *SHP-1*, *CYP7A1*, or β -actin.

To substantiate the *in vivo* data and extend them to human hepatocytes, we examined whether *SHP-1* and *CYP7A1* expression were regulated by FXR in primary cultures of rat and human hepatocytes. Hepatocytes were treated with increasing concentrations of GW4064, and the levels of *SHP-1* and *CYP7A1* expression were

examined by Northern blot analysis. GW4064 treatment markedly increased *SHP-1* expression and decreased *CYP7A1* expression in hepatocytes from both species in a dose-dependent fashion (Figure 2B). Similar results were obtained in human hepatocytes treated with the natural FXR ligand CDCA (Figure 2C). As expected, CDCA was less potent than GW4064 in its effects on gene expression (compare Figures 2B and 2C). These data strongly suggest that FXR regulates *SHP-1* and *CYP7A1* expression in both human and rodent hepatocytes. Notably, there was a striking reciprocal relationship between the regulation of *SHP-1* and *CYP7A1*

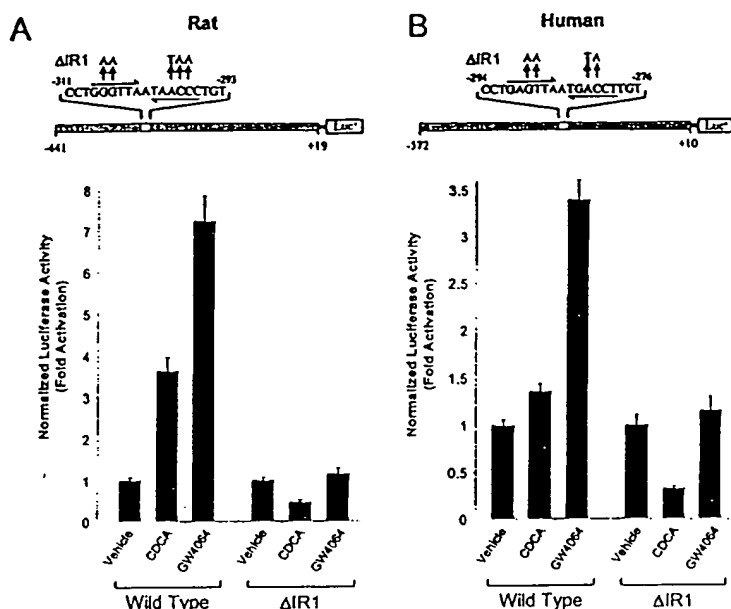


Figure 4. FXR Activates the Rat and Human SHP-1 Promoters

HepG2 cells were transfected with the human FXR expression plasmid and luciferase reporter plasmids containing the proximal promoters of the rat ([A], nucleotides -441 to +19) or human ([B], nucleotides -572 to +10) *SHP-1* genes or the corresponding reporter plasmids in which the IR1 elements had been mutated (Δ IR1). Following transfection, cells were treated for 48 hr with GW4064 (1 μ M) or CDCA (100 μ M). Data represent the mean \pm S.D. of six individual transfections.

expression: GW4064 and CDCA repressed *CYP7A1* expression at the same concentrations that were required to induce *SHP-1* expression (Figures 2B and 2C). Since *SHP-1* is known to heterodimerize with several other members of the nuclear receptor superfamily and to repress their transcriptional activity (Seol et al., 1996; Masuda et al., 1997; Johansson et al., 1999), these data raised the intriguing possibility that FXR-mediated induction of *SHP-1* might underlie the repression of *CYP7A1* expression (see below).

FXR Binds and Activates *SHP-1* Promoters

We next sought to determine whether *SHP-1* expression is directly regulated by FXR. FXR preferentially binds as a heterodimer with RXR to FXREs composed of two nuclear receptor half-sites of consensus AG(G/T)TCA organized as an inverted repeat and separated by a single nucleotide (IR1) (Forman et al., 1995). IR1-type FXREs have been identified in the human and mouse *I-BABP* promoters (Grober et al., 1999; Makishima et al., 1999). The mouse, rat, and human *SHP-1* promoters were examined for IR1 motifs. A highly conserved IR1-like element was identified ~300 nucleotides upstream of the transcription initiation site in the *SHP-1* promoter of all three species (Figure 3A). Electrophoretic mobility-shift analyses demonstrated that the FXR/RXR complex binds efficiently to the IR1 element from the *SHP-1* promoter of each species (Figure 3B). In agreement with earlier observations (Grober et al., 1999), the FXR/RXR heterodimer also bound to the mouse and human *I-BABP* FXREs (Figure 3B). Competition binding analyses showed that these interactions were specific: no competition was seen with a mutated derivative of the IR1 motif derived from the mouse *SHP-1* promoter (Figure 3C).

The presence of an FXR/RXR binding site suggested that the *SHP-1* gene is directly regulated by FXR. To test this hypothesis, HepG2 cells were transfected with an FXR expression plasmid and reporter plasmids expressing luciferase under the control of either the rat or

human *SHP-1* promoters. GW4064 treatment of cells transfected with the FXR expression plasmid and either promoter construct resulted in a marked induction of reporter activity (Figures 4A and 4B). Based on Northern blot analysis of *SHP-1* expression (Figure 2B), the magnitude of the response from the rat (7-fold) and human (3-fold) *SHP-1* promoters was somewhat lower than expected and it is possible that other promoter or enhancer elements contribute to the regulation of *SHP-1* expression. Alternately, additional factors present in well differentiated cultures of rat hepatocytes but not HepG2 cells may be required for maximal FXR responsiveness. In the absence of exogenously expressed FXR, the rat and human *SHP-1* promoters exhibited a modest (~1.5-fold) induction on exposure to GW4064, which is most likely due to endogenous FXR in HepG2 cells (data not shown). FXR responsiveness was eliminated when mutations were introduced into the IR1 motifs in either the rat or human *SHP-1* promoters (Figures 4A and 4B). These data provide strong evidence that *SHP-1* expression is regulated directly by the FXR/RXR heterodimer in multiple species.

SHP-1 Interacts with Orphan Nuclear Receptor LXR-1

The finding that *SHP-1* expression is regulated by FXR together with the reciprocal relationship between *SHP-1* and *CYP7A1* regulation (Figure 2) suggested that *SHP-1* might play a pivotal role in bile acid-mediated repression of *CYP7A1* expression. Regulation of the *CYP7A1* promoter is complex and involves numerous transcription factors, including nuclear receptors with known ligands such as the thyroid hormone receptor (TR), retinoic acid receptor (RAR), RXR and LXR α , and the orphan receptors COUP-TFII, HNF4 α , and LXR-1 (Lehmann et al., 1997; Stroup et al., 1997; Chiang, 1998; Peet et al., 1998; Nitta et al., 1999; Russell, 1999; Stroup and Chiang, 2000). *SHP-1* has previously been shown to bind to and repress the transcriptional activities of TR, RAR, and RXR in the presence of their ligands and HNF4 α in the

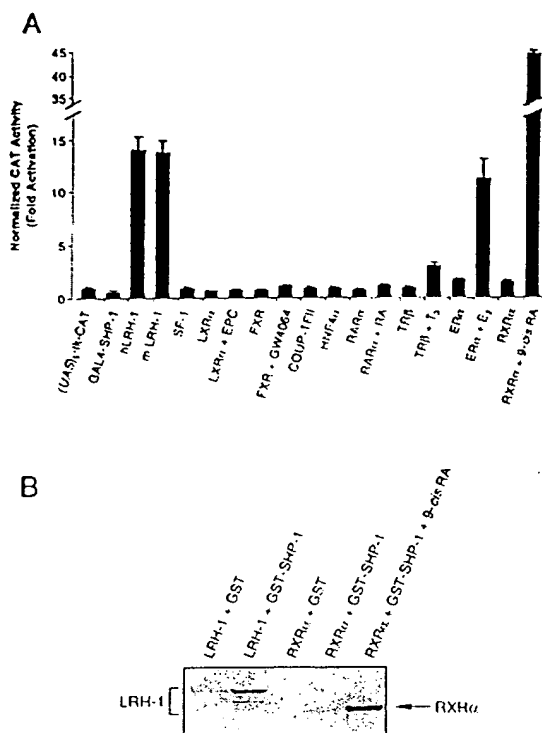


Figure 5. SHP-1 interacts with the Orphan Nuclear Receptor LRH-1. (A) Mammalian two-hybrid experiments were performed in CV-1 cells cotransfected with expression plasmids for the GAL4-human SHP-1 chimera and various VP16-nuclear receptor ligand-binding domain chimeras. Transfection assays containing the LXRα-, FXR-, RARα-, TRβ-, ERα-, and RXRα-GAL4 chimeras were performed in the absence or presence of the indicated ligands [respectively: EPC, 24(S),25-epoxycholesterol (10 μM), GW4064 (1 μM); RA, all-*trans* retinoic acid (0.1 μM); T₃, triiodothyronine (0.1 μM); E₂, estradiol (0.1 μM); 9-*cis* RA, 9-*cis* retinoic acid (0.1 μM)]. Data are expressed as fold activation over cells transfected with the (UAS)₃-tk-CAT reporter alone and represent the mean of assays (n = 8) ± S.D. (B) GST pull-down assays were performed with [³⁵S]-labeled LRH-1 or RXRα in the presence of GST or GST-SHP-1 as indicated. 9-*cis* retinoic acid (9-*cis* RA) was added to the binding reaction to a final concentration of 10 μM.

absence of any exogenous ligand (Seol et al., 1996; Masuda et al., 1997). Using a mammalian two-hybrid approach, we examined whether SHP-1 interacts with these and other nuclear receptors that have been implicated in the regulation of *CYP7A1*. CV-1 cells were transfected with an expression plasmid for a GAL4-SHP-1 chimera, the (UAS)₃-tk-CAT reporter, and expression plasmids for chimeras between the strong transcriptional activation domain of VP16 and the isolated ligand-binding domains of a panel of nuclear receptors (Figure 5A). When transfected alone, the GAL4-SHP-1 chimera caused a minor reduction (~0.3-fold) in reporter activity (Figure 5A). However, reporter activity was strongly induced when GAL4-SHP-1 was coexpressed with VP16-RXRα (~44-fold) or VP16-estrogen receptor α (ERα, ~11-fold) in the presence of 9-*cis* retinoic acid and estradiol, respectively (Figure 5A). These interactions were strongly dependent on the presence of ligand. Little or no interaction was detected between SHP-1 and LXRα,

FXR, COUP-TFII, HNF4α, RARα, or TRβ in our mammalian two-hybrid assay (Figure 5A). The lack of a stronger interaction between SHP-1 and either TRβ, RARα, or HNF4α was surprising in light of the previous results of others (Seol et al., 1996; Masuda et al., 1997) and may reflect differences in the assay systems used. Notably, strong reporter activity was detected when GAL4-SHP-1 was expressed with VP16-human LRH-1 or VP16-mouse LRH-1 (~14-fold activation for both human and mouse). This activity was completely dependent on the presence of GAL4-SHP-1 (data not shown). These data demonstrate that SHP-1 can interact with LRH-1 in cells. Interestingly, little or no interaction was detected between SHP-1 and steroidogenic factor 1 (SF-1) (Figure 5A), a closely related orphan receptor that shares ~60% amino acid identity with LRH-1 in the ligand-binding domain (Tsukiyama et al., 1992; Honda et al., 1993; Ikeda et al., 1993).

Using a glutathione S-transferase (GST) pull-down assay, we examined whether SHP-1 binds directly to LRH-1. SHP-1 was expressed in *E. coli* as a fusion protein with GST, and [³⁵S]-labeled LRH-1 was synthesized in vitro. Glutathione-Sepharose beads efficiently coprecipitated [³⁵S]-labeled LRH-1 in the presence of GST-SHP-1 but not in its absence (Figure 5B). In parallel incubations, GST-SHP-1 interacted strongly with [³⁵S]-labeled human RXRα in the presence of 9-*cis* retinoic acid (Figure 5B). These data are in close agreement with those derived from mammalian two-hybrid experiments (Figure 5A). Thus, SHP-1 interacts directly with LRH-1.

SHP-1 Represses Expression of *CYP7A1*

Does SHP-1 have a role in the repression of *CYP7A1* expression by FXR ligands? We addressed this question by performing cotransfection experiments with a rat *CYP7A1* luciferase reporter plasmid (pGL3-rCYP7A1 [-1573/+36]) containing nucleotides -1573 to +36 of the rat *CYP7A1* promoter, which includes a conserved LRH-1 binding site (Nitta et al., 1999). In the absence of exogenously expressed LRH-1, the activity of the pGL3-rCYP7A1 (-1573/+36) reporter was low when transiently transfected into HepG2 cells (data not shown). Cotransfection of increasing amounts of an LRH-1 expression plasmid resulted in a dose-dependent increase in reporter activity (Figure 6). This LRH-1-dependent reporter activity was completely blocked by the cotransfection of SHP-1 expression plasmid (Figure 6). These data suggest that interactions between SHP-1 and LRH-1 represent a basis for bile acid-mediated repression of *CYP7A1* expression.

Discussion

The recent discovery that FXR is a bile acid receptor provided a great deal of insight into the molecular mechanisms underlying bile acid signaling. In particular, these studies uncovered the mechanism whereby bile acids stimulate the transcription of genes, such as *I-BABP*, involved in bile acid transport. High-affinity binding sites for the FXR/RXR heterodimer have been identified in both the human and mouse *I-BABP* promoters (Grober et al., 1999; Makishima et al., 1999). By contrast, the mechanism underlying bile acid-mediated repression of *CYP7A1* expression remained a puzzle, since an FXRE had not been identified in the bile acid response elements of this gene (Chiang and Stroup, 1994; Chiang et

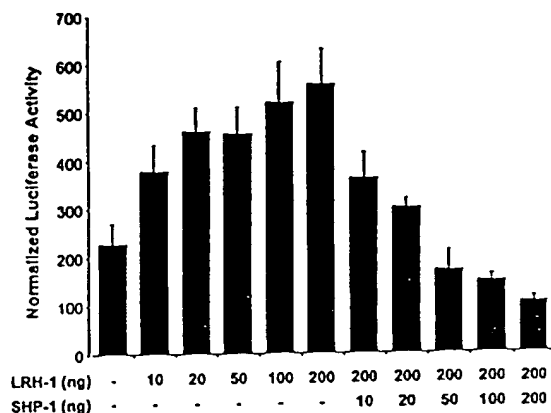


Figure 6. SHP-1 Represses LRH-1-Dependent Activation of the Rat *CYP7A1* Promoter

HepG2 cells were transfected with the rat *CYP7A1* reporter plasmid, pGL3-rCYP7A1(-1573/+36), and the indicated amounts of LRH-1 and/or SHP-1 expression plasmids. Data represent the mean of assays performed in triplicate \pm S.D.

al., 2000). We now present evidence that FXR does not repress *CYP7A1* expression directly, but rather through induction of the gene encoding the orphan nuclear receptor SHP-1, which, in turn, represses *CYP7A1* expression. Similar findings have been reported by Lu et al. (2000 [this issue of *Mol. Cell*]). Consistent with this model, it was recently shown that SHP-1 expression is markedly lower and not inducible by cholic acid in the livers of mice lacking FXR (Sinal et al., 2000). Taken together, these data provide a molecular explanation for the coordinate suppression of gene expression by bile acids.

SHP-1 Represses *CYP7A1* Expression

We encountered the orphan nuclear receptor SHP-1 as part of a comprehensive, unbiased effort to identify FXR target genes in the liver. SHP-1 expression was strongly induced in the livers of rats treated with the potent, nonsteroidal FXR ligand GW4064. SHP-1 expression was also markedly induced by GW4064 in primary cultures of human and rat hepatocytes, whereas *CYP7A1* expression was suppressed under the same conditions. The reciprocal relationship between SHP-1 and *CYP7A1* regulation, together with the established inhibitory effects of SHP-1 on nuclear receptor activity, suggested that SHP-1 might repress *CYP7A1* expression. Indeed, expression of SHP-1 repressed the activity of the rat *CYP7A1* promoter in HepG2 cells.

SHP-1 is unusual in that it lacks the highly conserved

DNA-binding domain typically found in members of the nuclear receptor family. SHP-1 was originally cloned in yeast two-hybrid experiments using the orphan nuclear receptors CAR or PPAR α as bait, but it interacts with a number of additional nuclear receptors, including ER α and ER β , RAR, RXR, and TR (Seol et al., 1996; Masuda et al., 1997; Seol et al., 1998; Johansson et al., 1999). In each case, SHP-1 represses the ligand-induced transcriptional activity of these receptors. How does SHP-1 repress transcription of the *CYP7A1* promoter? Our data indicate that SHP-1 exerts much of its effect through interaction with the orphan nuclear receptor LRH-1. SHP-1 interacted strongly with LRH-1 in both a mammalian two-hybrid assay and an in vitro pull-down assay. Moreover, SHP-1 efficiently repressed LRH-1-dependent activation of the rat *CYP7A1* promoter. LRH-1 was recently shown to activate the human *CYP7A1* promoter by binding to an extended nuclear receptor half-site sequence that is conserved in the mouse, rat, and hamster *CYP7A1* promoters (Nitta et al., 1999). Earlier studies had defined DNA response elements in the *CYP7A1* and *CYP8B1* gene promoters that conferred repression in response to bile acids (Chiang and Stroup, 1994; Chiang et al., 2000; del Castillo-Olivares and Gil, 2000). Notably, each of these negative bile acid response elements contains an LRH-1 binding site. Consistent with these data, *CYP8B1* expression was repressed 3-fold in Fisher rats treated with GW4064 (S. A. J., unpublished data). Thus, interactions between SHP-1 and LRH-1 are likely to be important for the coordinate repression of a number of genes by bile acids. Among the genes that may be regulated by the interaction between SHP-1 and LRH-1 is SHP-1 itself. An LRH-1-responsive region of the murine SHP-1 gene has been identified (Lee et al., 1999). Thus, SHP-1 is likely to regulate its own expression. This feedback regulation may provide a mechanism for attenuating the bile acid-mediated repression of genes by SHP-1. A model for bile acid-mediated repression of gene expression via increased SHP-1 levels is shown in Figure 7.

Two recent reports showed that SHP-1 represses the transcriptional activation of ER α and ER β , RXR, and the orphan receptor HNF4 α by competing with coactivator binding to these receptors (Johansson et al., 1999; Lee et al., 2000). In addition, SHP-1 contains a strong transcriptional repressor domain in its C terminus (Lee et al., 2000). Furthermore, SHP-1 has been shown to inhibit DNA binding of RAR-RXR heterodimers (Seol et al., 1996). Taken together, these studies suggest that SHP-1 inhibits the transcriptional activity of nuclear receptors through multiple mechanisms. To date, we have been unable to demonstrate inhibition of LRH-1 binding to its response element in the *CYP7A1* promoter by SHP-1 (data not shown). Thus, the mechanism by which SHP-1

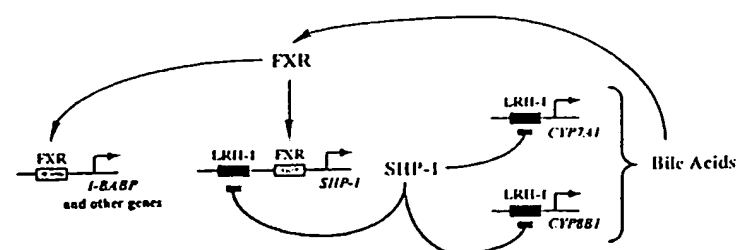


Figure 7. Model for the Feedforward and Feedback Regulatory Effects of Bile Acids on Gene Expression

Activation of FXR by bile acids results in the induction of I-BABP and SHP-1 expression. SHP-1, in turn, interacts with LRH-1 and represses expression of *CYP7A1* and *CYP8B1*. SHP-1 may also repress expression of its own gene.

Inhibits LRH-1-mediated transactivation of the *CYP7A1* promoter remains unresolved.

In addition to the interactions between SHP-1 and LRH-1, other mechanisms may play a role in bile acid-mediated repression of *CYP7A1* expression. First, SHP-1 binds to and represses the transcriptional activity of other nuclear receptors that regulate *CYP7A1*, including RXR and TR (Seol et al., 1996; Masuda et al., 1997). These interactions may also contribute to bile acid-mediated repression of *CYP7A1* expression. Second, ligand-bound FXR was reported to repress LXR α activity on an LXR α response element (Wang et al., 1999), although the mechanism for this *trans*-repression is not clear. Since LXR α stimulates rodent *CYP7A1* expression in response to oxysterols, repression of LXR α activity may contribute to the overall repression of *CYP7A1*. Thus, SHP-1/LRH-1 interactions may be one of several mechanisms whereby bile acids repress expression of *CYP7A1* and other genes.

Parallels between SHP-1/LRH-1 and Other Nuclear Receptor Pairs

Intriguing parallels exist between the SHP-1/LRH-1 interaction and another pair of nuclear receptors. LRH-1 is most closely related to the orphan receptor SF-1, which regulates the expression of enzymes required for steroid hormone biosynthesis (Parker, 1998; Hammer and Ingraham, 1999). SF-1 and LRH-1 are ~85% identical in the amino acid sequences of their DNA-binding domains, and both bind as monomers to the same extended nuclear receptor half-site sequence. Notably, the transcriptional activity of SF-1 is repressed by binding to DAX-1 (dosage-sensitive sex-reversal adrenal hypoplasia congenital region on the X chromosome, region 1; NR0B1), an orphan nuclear receptor most closely related to SHP-1 that also lacks the DNA-binding domain characteristic of nuclear receptors (Zanaria et al., 1994; Hammer and Ingraham, 1999). Thus, both SF-1 and LRH-1 are negatively regulated in a *trans*-dominant fashion by heterodimerization with orphan receptors lacking DNA-binding domains. Since SHP-1 expression is stimulated by bile acids, it will be interesting to determine whether DAX-1 expression is also regulated by hormones.

A second nuclear receptor pair with similarities to SHP-1/LRH-1 occurs in *Drosophila*. Hormonal activation of the ecdysone receptor (EcR) during the third larval instar phase of *Drosophila* metamorphosis results in an increase in the expression of two orphan nuclear receptors, DHR3, which has a functional DNA-binding domain, and E75B, which does not. E75B binds to DHR3 and represses its transcriptional activity (Thummel, 1997; White et al., 1997). This interaction is critical for determining the temporal progression of metamorphosis. The EcR/E75/DHR3 and FXR/SHP-1/LRH-1 regulatory cascades are remarkably similar in that hormone-mediated activation of a nuclear receptor (either FXR or EcR) induces expression of a second nuclear receptor, which, in turn, binds to and represses the activity of a third nuclear receptor. The similarities in these genetic hierarchies across evolution suggest that repression via heterodimerization may represent an important paradigm for the modulation of orphan receptor activity.

Conclusions

The mechanism whereby FXR represses expression of *CYP7A1* and other genes has until now remained an

enigma. Through the use of a potent, nonsteroidal FXR ligand, we have identified *SHP-1* as an FXR target gene in the liver of humans and rodents. Furthermore, we have demonstrated that SHP-1 can interact with LRH-1 and efficiently repress expression of *CYP7A1*. Thus, bile acid-induced repression of *CYP7A1* is mediated by a novel regulatory cascade of three nuclear receptors. Since both the *CYP7A1* and *CYP8B1* gene promoters contain LRH-1 binding sites, the SHP-1/LRH-1 partnership is likely to have broad implications in bile acid signaling. Both SHP-1 and LRH-1 are orphan receptors, which raises the possibility that bile acid biosynthesis will be regulated by additional, unidentified hormones. Regardless of whether SHP-1 and LRH-1 have natural ligands, pharmacologic modulation of their interaction represents an exciting new opportunity for the discovery of drugs that regulate cholesterol homeostasis.

Experimental Procedures

Materials

The synthesis of GW4064 will be described elsewhere (Maloney et al., 2000). CDCA, dexamethasone, estradiol, all-*trans* retinoic acid, 9-*cis* retinoic acid, and charcoal-stripped, delipidated calf serum were acquired from the Sigma Chemical Co. (St. Louis, MO). 24(S),25-epoxycholesterol was synthesized in-house. DNA-modifying enzymes, polymerases, and restriction endonucleases were provided by Roche Molecular Biochemicals (Indianapolis, IN). Charcoal/dextran-treated fetal bovine serum (FBS) was purchased from Hyclone Laboratories Inc. (Logan, UT). The human hepatocellular carcinoma cell line HepG2 was obtained from the American Type Culture Collection (ATCC number HB-8065, Manassas, VA). Matrigel was provided by Becton Dickinson Labware (Bedford, MA). All other tissue culture reagents were obtained from Life Technologies Inc. (Gaithersburg, MD).

Animals

Male Fisher rats were obtained from Charles River Laboratories Inc. (Raleigh, NC) and maintained on a 12 hr light/12 hr dark cycle. Animals were allowed food and chow ad libitum. GW4064 (30 mg/kg) was administered by gavage twice a day for 7 days and the animals sacrificed by cervical dislocation 4 hr after the final treatment. Livers were excised and snap-frozen in liquid nitrogen. Differential gene expression analysis was performed by CuraGen Corp. (New Haven, CT).

Plasmid Constructs

Expression plasmids for the human nuclear receptor-GAL4 chimeras were prepared by inserting amplified cDNAs encoding the ligand-binding domains into a modified pSG5 expression vector (Stratagene, La Jolla, CA) containing the GAL4 DNA-binding domain (amino acids 1-147) and the Simian virus 40 (SV40) large T antigen nuclear localization signal (APKKKRRKVG). The (UAS)₃-tk-CAT and (hsp27EcRE)₃-tk-LUC reporter constructs have been previously described (Forman et al., 1995; Parks et al., 1999). p β -actin-SPAP, an expression vector containing the human secreted placental alkaline phosphatase (SPAP) cDNA under the control of β -actin promoter, was used as an internal control in all transfections. The expression plasmids for human and mouse FXR (pSG5-hFXR and pSG5-mFXR, respectively) and human SRC-1 are described elsewhere (Kliwer et al., 1998; Parks et al., 1999). The full-length coding regions for human LRH-1 (GenBank Accession Number AB019246) and human SHP-1 (GenBank Accession Number L76571) were amplified by PCR and cloned into pSG5, creating pSG5-hLRH-1 and pSG5-hSHP-1, respectively. A consensus Kozak sequence was created during amplification. The rat (bases -441 to +19, GenBank Accession Number D86745) (Masuda et al., 1997) and human (bases -572 to +10, GenBank Accession Number AF044316) (Lee et al., 1998) *SHP-1* promoters were amplified by PCR using the following primer pairs: Rat, 5'-gggtgtgcgagatctCCTGGCTGGCTCCTGGCTCTGT-3' (sense) and 5'-gggtgtgcgagatctCCTGTTCTTCTGGCTCTGT

GGC-3' (antisense); and human, 5'-gggtgtgagatctCTAGACT GGACAGTGGGCAAAG-3' (sense) and 5'-gggtgtgagatctCTCC AGCTCTCTGGCTCTGTGT-3' (antisense). The resultant fragments were inserted into the *Bgl*II site of pGL3-Basic, a promoter-less luciferase reporter vector (Promega, Madison, WI). Site-directed mutagenesis of putative FXREs in the rat and human *SHP-1* promoters was performed using the Transformer mutagenesis system (CLONTECH Laboratories, Palo Alto, CA) with the *Δ*ratIR1 (bases -321 to -287, 5'-CCTGGTACAGCCTGGaaTAATaCTGTTTATAC-3') and *Δ*humanIR1 (bases -304 to -270, 5'-CCTGGTACAGCCTGA aaTAATGtaCTGTTTATCC-3') primers. Mutated constructs were verified to be free of nonspecific base changes by sequencing. pGL3-CYP7A1(-1573/+36) contains bases -1573 to +36 of the rat *CYP7A1* promoter (GenBank Accession Number Z14108) inserted into the *Nhe*I site of pGL3-Basic. VP16-nuclear receptor chimeras contain the 80 aa Herpes virus VP16 transactivation domain linked to the ligand-binding domain of the following nuclear receptors in a modified pSG5 expression vector: human COUP-TFII, ER α , LHR-1, LXR α , RAR α , and TR β ; mouse FXR, LHR-1, RXR α , and SF-1; and rat HNF4 α .

Transient Transfection Assays

Transient transfection of CV-1 cells was performed exactly as described elsewhere (Jones et al., 2000). Typically, transfection mixes contained 2–5 ng of receptor expression vector, 20 ng of reporter construct, and 8 ng of p β -actin-SPAP. The amount of DNA used in each transfection was adjusted to 80 ng with carrier plasmid (pBluescript, Stratagene). Mammalian two-hybrid experiments utilized transfection mixes containing 20 ng of VP16 nuclear receptor ligand-binding domain expression vector, 5 ng of pSG5-GAL4-SHP-1, 15 ng of (UAS)₃-tk-CAT, and 8 ng of p β -actin-SPAP. Cells were maintained for 24 hr in the presence of drug (added as a 1000 \times stock in dimethyl sulfoxide) in DMEM/F-12 nutrient mixture containing 10% charcoal-stripped, delipidated calf serum. An aliquot of medium was assayed for SPAP activity, and the cells were lysed prior to determination of luciferase expression. Luciferase activities were normalized to SPAP. HepG2 cells were maintained in DMEM/F-12 supplemented with 10% heat-inactivated FBS (Life Technologies Inc.). Plasmid DNA was transfected into HepG2 cells using the FuGENE6 transfection reagent according to the manufacturer's instructions (Roche Molecular Biochemicals). Thus, 24-well culture plates (15 mm diameter) were inoculated with 7×10^5 cells 24 hr prior to transfection. Cells were transfected overnight in serum-free DMEM/F-12 with 100 ng of reporter construct, 32 ng of p β -actin-SPAP, and 0–400 ng of receptor expression vectors (adjusted to 400 ng with carrier plasmid). Following transfection, the medium was aspirated and the cells were cultured for a further 48 hr in DMEM/F-12 supplemented with 10% heat-inactivated FBS. SPAP and luciferase values were determined as described above.

Primary Culture of Human and Rat Hepatocytes and Northern Blot Analysis

Primary human hepatocytes were obtained from Dr. Steve Strom (University of Pittsburgh). Rat hepatocytes were isolated as described elsewhere (LeCluyse et al., 1996). Cells (1.5×10^6) were cultured on Matrigel-coated 6-well plates in serum-free Williams' E medium supplemented with 100 nM dexamethasone, 100 U/ml penicillin G, 100 μ g/ml streptomycin, and insulin-transferrin-selenium (ITS-G, Life Technologies Inc.). Twenty-four hours after isolation, hepatocytes were treated with either GW4064 (0.1–10 μ M) or CDCA (1–100 μ M), which were added to the culture medium as 1000 \times stocks in dimethyl sulfoxide. Control cultures received vehicle alone. Cells were cultured for a further 48 hr prior to harvest, and total RNA was isolated using a commercially available reagent (Trizol, Life Technologies Inc.) according to the manufacturer's instructions. Total RNA (10 μ g) was resolved on a 1% agarose/2.2 M formaldehyde denaturing gel and transferred to a nylon membrane (Hybond N+, Amersham Pharmacia Biotech Inc., Piscataway, NJ). Blots were hybridized with ³²P-labeled cDNAs corresponding to human *SHP-1* (GenBank Accession Number L76571), human *CYP7A1* (bases 89–1564, GenBank Accession Number M93133), mouse *SHP-1* (bases 30–783, GenBank Accession Number L76567), or rat *CYP7A1* (bases 235–460, GenBank Accession Number J05460).

Subsequently, blots were stripped and reprobed with a radiolabeled β -actin cDNA (CLONTECH Laboratories).

Electrophoretic Mobility-Shift Assays

Electrophoretic mobility-shift assays (EMSA) were performed essentially as described elsewhere (Lehmann et al., 1997). hFXR and hRXR α were synthesized from pSG5-hFXR and pSG5-hRXR α expression vectors, respectively, using the TNT T7 Coupled Reticulocyte System (Promega). Unprogrammed lysate was prepared using the pSG5 expression vector (Stratagene). Binding reactions contained 10 mM HEPES (pH 7.8), 60 mM KCl, 0.2% Nonidet P-40, 6% glycerol, 2 mM dithiothreitol (DTT), 2 μ g of poly(dI-dC)·poly(dI-dC), and 1 μ l each of synthesized hFXR or hRXR α . Control incubations received unprogrammed lysate alone. Reactions were preincubated on ice for 10 min prior to the addition of [³²P]-labeled double-stranded oligonucleotide probe (0.2 pmol). Competitor oligonucleotides were added to the preincubation at 5-, 25-, and 75-fold molar excess. Samples were held on ice for a further 20 min, and the protein-DNA complexes resolved on a pre-electrophoresed 5% polyacrylamide gel in 0.5 \times TBE (45 mM Tris-borate, 1 mM EDTA) at room temperature. Gels were dried and autoradiographed at -70°C for 1–2 hr. The following double-stranded oligonucleotides were used as probes and competitors in EMSA: rSHP, 5'-gatcCCTG GGTTAATAACCCCTGT-3'; mSHP, 5'- gatcCCTGGGTTAATGACCC TGT-3'; hSHP, 5'- gatcCCTGAGTTAATGACCTTGT-3'; ml-BABP, 5'-gatcTTAAGGTGAATAACCTTGG-3'; hl-BABP, 5'-gatcCCAGGT GAATAACCTCGG-3' (Grober et al., 1999); and mSHPmut 5'-gatcCC TGGaaTAATGttCCTGT-3'.

GST Pull-Down Assays

GST-SHP-1 fusion protein was expressed in BL21(DE3)plysS cells and bacterial extracts prepared by one cycle of freeze-thaw of the cells in protein lysis buffer containing 50 mM Tris (pH 8.0), 250 mM KCl, 1% Triton X-100, 10 mM DTT and 1 \times Complete Protease Inhibitor (Roche Molecular Biochemical) followed by centrifugation at 40,000 \times g for 30 min. Glycerol was added to the resultant supernatant to a final concentration of 10%. Lysates were stored at -80°C until use. [³⁵S]-labeled human LHR-1 or human RXR α was generated using TNT T7 Coupled Reticulocyte System (Promega) in the presence of Pro-Mix (Amersham Pharmacia Biotech Inc.). Coprecipitation reactions included 25 μ l of lysate containing GST-SHP-1 fusion protein or control GST; 25 μ l of incubation buffer (50 mM KCl, 40 mM HEPES [pH 7.5], 5 mM β -mercaptoethanol, 0.1% Tween 20 and 1% nonfat dry milk); and 5 μ l of [³⁵S]-labeled LHR-1 or RXR α . The mixtures were incubated for 25 min with gentle rocking at 4°C prior to the addition of 20 μ l of glutathione-Sepharose 4B beads (Amersham Pharmacia Biotech Inc.) that had been extensively washed in protein lysis buffer. Reactions were incubated at 4°C with gentle rocking for a further 20 min. The beads were pelleted at 3000 rpm in a microfuge and washed four times with protein incubation buffer. Following the final wash, the beads were resuspended in 25 μ l of 2 \times SDS-PAGE sample buffer containing 50 mM DTT. Samples were heated to 100°C for 5 min and resolved on a 10% acrylamide gel. Autoradiography was performed overnight.

Statistical Analyses

Unless otherwise stated, data are expressed as mean \pm standard deviation (S.D.). The significance of differences in *SHP-1* and *CYP7A1* expression between vehicle- and GW4064-treated animals were analyzed using an unpaired Student's *t*-test.

Acknowledgments

We thank Dr. Traci Mansfield (CuraGen Corp., New Haven, CT) for assistance with the CuraGen data analysis, Dr. Geraldine Hamilton (University of North Carolina, Chapel Hill) for preparation of the rat hepatocytes, James Way for advice on statistical analyses, Dr. Scott Sundseth for providing the rat *CYP7A1* cDNA, and Drs. Rich Buckholz and Catherine Stoltz for comments on the manuscript.

Received May 23, 2000; revised July 18, 2000.

References

- Apfel, R., Benbrook, D., Lemhardt, E., Ortiz, M.A., Salbert, G., and Pfahl, M. (1994). A novel orphan receptor specific for a subset of thyroid hormone-responsive elements and its interaction with the retinoid/thyroid hormone receptor subfamily. *Mol. Cell. Biol.* 14, 7025-7035.
- Becker-Andre, M., Andre, E., and DeLamarier, J.F. (1993). Identification of nuclear receptor mRNAs by RT-PCR amplification of conserved zinc-finger motif sequences. *Biochem. Biophys. Res. Commun.* 194, 1371-1379.
- Broadus, J., McCabe, J.R., Endrizzi, B., Thummel, C.S., and Woodward, C.T. (1999). The *Drosophila* β FTZ-F1 orphan nuclear receptor provides competence for stage-specific responses to the steroid hormone ecdysone. *Mol. Cell* 3, 143-149.
- Chiang, J.Y., and Stroup, D. (1994). Identification and characterization of a putative bile acid-responsive element in cholesterol 7 α -hydroxylase gene promoter. *J. Biol. Chem.* 269, 17502-17507.
- Chiang, J.Y.L. (1998). Regulation of bile acid synthesis. *Front. Biosci.* 3, 176-193.
- Chiang, J.Y.L., Kimmel, R., Weinberger, C., and Stroup, D. (2000). Farnesoid X receptor responds to bile acids and represses cholesterol 7 α -hydroxylase gene (CYP7A1) transcription. *J. Biol. Chem.* 275, 10918-10924.
- del Castillo-Olivares, A., and Gil, G. (2000). α -Fetoprotein transcription factor is required for the expression of sterol 12 α -hydroxylase, the specific enzyme for cholic acid synthesis. *J. Biol. Chem.*, 275, 17793-17799.
- Forman, B.M., Goode, E., Chen, J., Oro, A.E., Bradley, D.J., Perlmann, T., Noonan, D.J., Burkha, L.T., McMorris, T., Lamph, W.W., et al. (1995). Identification of a nuclear receptor that is activated by farnesol metabolites. *Cell* 81, 687-693.
- Galarneau, L., Pare, J.F., Allard, D., Hamel, D., Levesque, L., Tugwood, J.D., Green, S., and Belanger, L. (1996). The alpha1-fetoprotein locus is activated by a nuclear receptor of the *Drosophila* FTZ-F1 family. *Mol. Cell. Biol.* 16, 3853-3865.
- Grober, J., Zaghini, I., Fujii, H., Jones, S.A., Kliewer, S.A., Willson, T.M., Ono, T., and Besnard, P. (1999). Identification of a bile acid-responsive element in the human ileal bile acid-binding protein gene. Involvement of the farnesoid X receptor/9-cis-retinoic acid receptor heterodimer. *J. Biol. Chem.* 274, 29749-29754.
- Gustafsson, J.A. (1999). Seeking ligands for lonely orphan receptors. *Science* 284, 1285-1286.
- Hammer, G.D., and Ingraham, H.A. (1999). Steroidogenic factor-1: its role in endocrine organ development and differentiation. *Front. Neuroendocrinol.* 20, 199-223.
- Honda, S., Morohashi, K., Nomura, M., Takeya, H., Kitajima, M., and Omura, T. (1993). Ad4BP regulating steroidogenic P-450 gene is a member of steroid hormone receptor superfamily. *J. Biol. Chem.* 268, 7494-7502.
- Ikeda, Y., Lala, D.S., Luo, X., Kim, E., Moisan, M.P., and Parker, K.L. (1993). Characterization of the mouse FTZ-F1 gene, which encodes a key regulator of steroid hydroxylase gene expression. *Mol. Endocrinol.* 7, 852-860.
- Javitt, N.B. (1994). Bile acid synthesis from cholesterol: regulatory and auxiliary pathways. *FASEB J.* 8, 1308-1311.
- Johansson, L., Thomsen, J.S., Dandimopoulos, A.E., Spyrou, G., Gustafsson, J.A., and Treuter, E. (1999). The orphan nuclear receptor SHP inhibits agonist-dependent transcriptional activity of estrogen receptors ER α and ER β . *J. Biol. Chem.* 274, 345-353.
- Jones, S.A., Moore, L.B., Shenk, J.L., Wisely, G.B., Hamilton, G.A., McKee, D.D., Tomkinson, N.C., LeCluyse, E.L., Lambert, M.H., Willson, T.M., et al. (2000). The pregnane X receptor: a promiscuous xenobiotic receptor that has diverged during evolution. *Mol. Endocrinol.* 14, 27-39.
- Kliewer, S.A., Moore, J.T., Wade, L., Staudinger, J.L., Watson, M.A., Jones, S.A., McKee, D.D., Oliver, B.B., Willson, T.M., Zetterstrom, R.H., et al. (1998). An orphan nuclear receptor activated by pregnanes defines a novel steroid signaling pathway. *Cell* 82, 73-82.
- Lavorgna, G., Ueda, H., Cios, J., and Wu, C. (1991). FTZ-F1, a steroid hormone receptor-like protein implicated in the activation of fushi tarazu. *Science* 252, 848-851.
- LeCluyse, E., Bullock, P., Parkinson, A., and Hochman, J.H. (1996). Cultured rat hepatocytes. In *Model Systems for Biopharmaceutical Assessment of Drug Absorption and Metabolism*, R.T. Borchardt, G. Wilson, and P. Smith, eds. (New York: Plenum), pp. 121-159.
- Lee, H.K., Lee, Y.K., Park, S.H., Kim, Y.S., Lee, J.W., Kwon, H.B., Soh, J., Moore, D.D., and Choi, H.S. (1998). Structure and expression of the orphan nuclear receptor SHP gene. *J. Biol. Chem.* 273, 14398-14402.
- Lee, Y.K., Dell, H., Dowhan, D.H., Hadzopoulou-Cladaras, M., and Moore, D.D. (2000). The orphan nuclear receptor SHP inhibits hepatocyte nuclear factor 4 and retinoid X receptor transactivation: two mechanisms for repression. *Mol. Cell. Biol.* 20, 187-195.
- Lee, Y.K., Parker, K.L., Choi, H.S., and Moore, D.D. (1999). Activation of the promoter of the orphan receptor SHP by orphan receptors that bind DNA as monomers. *J. Biol. Chem.* 274, 20869-20873.
- Lehmann, J.M., Kliewer, S.A., Moore, L.B., Smith-Oliver, T.A., Oliver, B.B., Su, J.L., Sundseth, S.S., Winegar, D.A., Blanchard, D.E., Spencer, T.A., and Willson, T.M. (1997). Activation of the nuclear receptor LXR by oxysterols defines a new hormone response pathway. *J. Biol. Chem.* 272, 3137-3140.
- Lehmann, J.M., McKee, D.D., Watson, M.A., Willson, T.M., Moore, J.T., and Kliewer, S.A. (1997). The human orphan nuclear receptor PXR is activated by compounds that regulate CYP3A4 gene expression and cause drug interactions. *J. Clin. Invest.* 102, 1016-1023.
- Lu, T.T., Makishima, M., Repa, J.J., Schoonjans, K., Kerr, T.A., Auwerx, J., and Mangelsdorf, D.J. (2000). Molecular basis for feedback regulation of bile acid synthesis by nuclear receptors. *Mol. Cell* 6, this issue, 507-515.
- Makishima, M., Okamoto, A.Y., Repa, J.J., Tu, H., Learned, R.M., Luk, A., Hull, M.V., Lustig, K.D., Mangelsdorf, D.J., and Shan, B. (1999). Identification of a nuclear receptor for bile acids. *Science* 284, 1362-1365.
- Maloney, P.R., Parks, D.J., Haffner, C.D., Fivush, A.M., Chandra, G., Plunket, K.D., Creech, K.L., Moore, L.B., Wilson, J.G., Lewis, M.C., et al. (2000). Identification of a chemical tool for the orphan nuclear receptor FXR. *J. Med. Chem.* 43, 2971-2974.
- Mangelsdorf, D.J., and Evans, R.M. (1995). The RXR heterodimers and orphan receptors. *Cell* 83, 841-850.
- Masuda, N., Yasuno, H., Tamura, T., Hashiguchi, N., Furusawa, T., Tsukamoto, T., Sadano, H., and Osumi, T. (1997). An orphan nuclear receptor lacking a zinc-finger DNA-binding domain: interaction with several nuclear receptors. *Biochim. Biophys. Acta* 1, 27-32.
- Nitta, M., Ku, S., Brown, C., Okamoto, A.Y., and Shan, B. (1999). CPF: an orphan nuclear receptor that regulates liver-specific expression of the human cholesterol 7 α -hydroxylase gene. *Proc. Natl. Acad. Sci. USA* 96, 6660-6665.
- Parker, K.L. (1998). The roles of steroidogenic factor 1 in endocrine development and function. *Mol. Cell. Endocrinol.* 145, 15-20.
- Parks, D.J., Blanchard, S.G., Bledsoe, R.K., Chandra, G., Consler, T.G., Kliewer, S.A., Stimmel, J.B., Willson, T.M., Zavacki, A.M., Moore, D.D., and Lehmann, J.M. (1999). Bile acids: natural ligands for an orphan nuclear receptor. *Science* 284, 1365-1368.
- Peet, D.J., Janowski, B.A., and Mangelsdorf, D.J. (1998). The LXRs: a new class of oxysterol receptors. *Curr. Opin. Genet. Dev.* 8, 571-575.
- Peet, D.J., Turley, S.D., Ma, W., Janowski, B.A., Lobaccaro, J.M., Hammer, R.E., and Mangelsdorf, D.J. (1998). Cholesterol and bile acid metabolism are impaired in mice lacking the nuclear oxysterol receptor LXR α . *Cell* 93, 693-704.
- Russell, D.W. (1999). Nuclear orphan receptors control cholesterol catabolism. *Cell* 97, 539-542.
- Russell, D.W., and Setchell, K.D. (1992). Bile acid biosynthesis. *Biochemistry* 31, 4737-4749.
- Seol, W., Choi, H.S., and Moore, D.D. (1995). Isolation of proteins that interact specifically with the retinoid X receptor: two novel orphan receptors. *Mol. Endocrinol.* 9, 72-85.
- Seol, W., Choi, H.S., and Moore, D.D. (1996). An orphan nuclear

hormone receptor that lacks a DNA binding domain and heterodimerizes with other receptors. *Science* 272, 1336–1339.

Seol, W., Hanstein, B., Brown, M., and Moore, D.D. (1998). Inhibition of estrogen receptor action by the orphan receptor SHP (short heterodimer partner). *Mol. Endocrinol.* 12, 1551–1557.

Shinkets, R.A., Lowe, D.G., Tai, J.T., Sehl, P., Jin, H., Yang, R., Predki, P.F., Rothberg, B.E., Murtha, M.T., Roth, M.E., et al. (1999). Gene expression analysis by transcript profiling coupled to a gene database query. *Nat. Biotechnol.* 17, 798–803.

Sinal, C.J., Tohkin, M., Miyata, M., Ward, J.M., Lambert, G., and Gonzalez, F.J. (2000). Targeted disruption of the nuclear receptor FXR/BAR impairs bile acid and lipid homeostasis. *Cell* 102, 731–744.

Stroup, D., and Chiang, J.Y. (2000). HNF4 and COUP-TFII interact to modulate transcription of the cholesterol 7 α -hydroxylase gene (CYP7A1). *J. Lipid Res.* 41, 1–11.

Stroup, D., Crestani, M., and Chiang, J.Y. (1997). Orphan receptors chicken ovalbumin upstream promoter transcription factor II (COUP-TFII) and retinoid X receptor (RXR) activate and bind the rat cholesterol 7 α -hydroxylase gene (CYP7A). *J. Biol. Chem.* 272, 9833–9839.

Thummel, C.S. (1997). Dueling orphans—interacting nuclear receptors coordinate *Drosophila* metamorphosis. *Bioessays* 19, 669–672.

Tsukiyama, T., Ueda, H., Hirose, S., and Niwa, O. (1992). Embryonal long terminal repeat-binding protein is a murine homolog of FTZ-F1, a member of the steroid receptor superfamily. *Mol. Cell. Biol.* 12, 1286–1291.

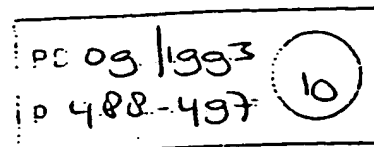
Wang, D.P., Stroup, D., Marrapodi, M., Crestani, M., Galli, G., and Chiang, J.Y. (1996). Transcriptional regulation of the human cholesterol 7 α -hydroxylase gene (CYP7A) in HepG2 cells. *J. Lipid Res.* 37, 1831–1841.

Wang, H., Chen, J., Hollister, K., Sowers, L.C., and Forman, B.M. (1999). Endogenous bile acids are ligands for the nuclear receptor FXR/BAR. *Mol. Cell* 3, 543–553.

White, K.P., Hurban, P., Watanabe, T., and Hogness, D.S. (1997). Coordination of *Drosophila* metamorphosis by two ecdysone-induced nuclear receptors. *Science* 276, 114–117.

Willy, P.J., Umesono, K., Ong, E.S., Evans, R.M., Heyman, R.A., and Mangelsdorf, D.J. (1995). LXR, a nuclear receptor that defines a distinct retinoid response pathway. *Genes Dev.* 9, 1033–1045.

Zanaria, E., Muscatelli, F., Bardoni, B., Strom, T.M., Guioli, S., Guo, W., Lalli, E., Moser, C., Walker, A.P., McCabe, E.R., et al. (1994). An unusual member of the nuclear hormone receptor superfamily responsible for X-linked adrenal hypoplasia congenita. *Nature* 372, 635–641.



The rapid proliferation and identification of newly cloned GPCRs reveal a much greater diversity within this supergene family than was previously considered at the pharmacological level.

Molecular Biology of G-Protein-Coupled Receptors

by Norman H. Lee
and Anthony R. Kerlavage

THE TRANSFER OF INFORMATION across the cell plasma membrane is a critical feature for the proper functioning of living cells. For many hormones, neurotransmitters and chemotactic factors, signal transduction is accomplished through the specific interaction of these bioactive molecules (agonists) with cell-surface receptors that couple to guanine nucleotide-binding regulatory proteins (G-proteins) (for a review see reference 1). The consequence of receptor occupancy by agonist is the generation of an intracellular second messenger signal that causes the cell to respond in an appropriate manner. G-protein-coupled receptors (GPCRs) play a key role in many physiologic processes, including nerve-to-nerve transmission, cardiac and smooth muscle contraction/relaxation, endocrine and exocrine secretion and chemotaxis. The fact that GPCRs mediate a broad spectrum of cellular events make these proteins an

ideal target for drug interaction and therapeutics.

As with all members of the GPCR gene family, the mechanism of signal transduction involves receptor coupling to a G-protein (for reviews see references 2 and 3). G-proteins are heterotrimeric proteins formed of a single GDP-bound α -subunit, one β -subunit and one γ -subunit. In response to agonist binding, GPCRs undergo a change in conformation (receptor-activated state) that triggers the formation of an agonist/receptor/G-protein ternary complex. Concomitant to ternary complex formation is the exchange of GDP for GTP on the α -subunit, thereby freeing the α -subunit from the $\beta\gamma$ -subunits. Consequently, the GTP-containing α -subunit (and in some cases the $\beta\gamma$ -subunits) acts to stimulate or inhibit an array of effector enzymes including adenylyl and guanylyl cyclase, phospholipase A and C, phosphodiesterases and ion channels. Termination of the signal transduction cascade is accomplished by the intrinsic GTPase activity found on the α -subunit. Hydrolysis of

bound GTP to GDP and inorganic phosphate leads to reassociation of the α -subunit with the $\beta\gamma$ -subunits and dissociation of the agonist/receptor/G-protein complex.

The first member of the GPCR gene family whose sequence was elucidated was the visual photoreceptor rhodopsin.^{4,5} During the past ten years, the number of cloned receptors has steadily risen and now approaches 200.⁶ These proteins are single polypeptides ranging in size from about 400–1000 amino acids. The activating ligand for GPCRs varies widely in character (Table I), yet these receptors share a highly conserved structure and topography. The hallmark feature of GPCRs is the presence of seven relatively hydrophobic domains, each 20–28 amino acids in length, that are presumed to span the lipid bilayer in an α -helical arrangement (Fig. 1). For the most part, it is the membrane-spanning regions which exhibit the greatest degree of amino acid sequence identity, ranging from 20% to more than 50%, depending on which receptor proteins are being compared.⁷ More divergent

**TABLE I: ENDOGENOUS LIGANDS
FOR G-PROTEIN-COUPLED
RECEPTORS**

Biogenics amines/neurotransmitters
Acetylcholine
Adenosine
Dopamine
Epinephrine
Glutamate
Histamine
Norepinephrine
Octopamine
Serotonin
Peptides/peptide hormones
Angiotensin
Bombesin-like peptides (neuromedin B, gastrin-releasing peptide)
Bradykinin
CSa anaphylatoxin
Calcitonin
Endothelin
N-formyl peptide
Interleukin-8
Neuromedin K (also known as neurokinin B)
Neuropeptide Y
Neurotensin
Parathyroid hormone/parathyroid related peptides
Secretin
Somatostatin
Substance K (also known as neurokinin A)
Substance P
Thyrotropin-releasing hormone
Vasoactive intestinal peptide
Glycoprotein hormones
Follicle-stimulating hormone
Lutropin/choriogonadotropin
Thyroid-stimulating hormone (also known as thyrotropin)
Regulatory factors
cAMP
Cannabinoids
Platelet-activating factor
Thromboxane A ₂
Thrombin
Yeast-mating factors (α and α -phero- mones)
Miscellaneous
Light
Odorants

are the extracellular amino- and intra-cellular carboxyl-terminal regions, as well as the six hydrophilic regions that connect the hydrophobic domains of the receptor to form alternating extracellular (e1, e2, e3) and intracellular (i1, i2, i3) loops (Fig. 1). This current model for the tertiary structure of GPCRs is based on analogy with bacteriorhodopsin, a light-activated proton pump whose three-dimensional structure was deduced from electron microscopy.^{8,9} The structure of bacteriorhodopsin is seen as

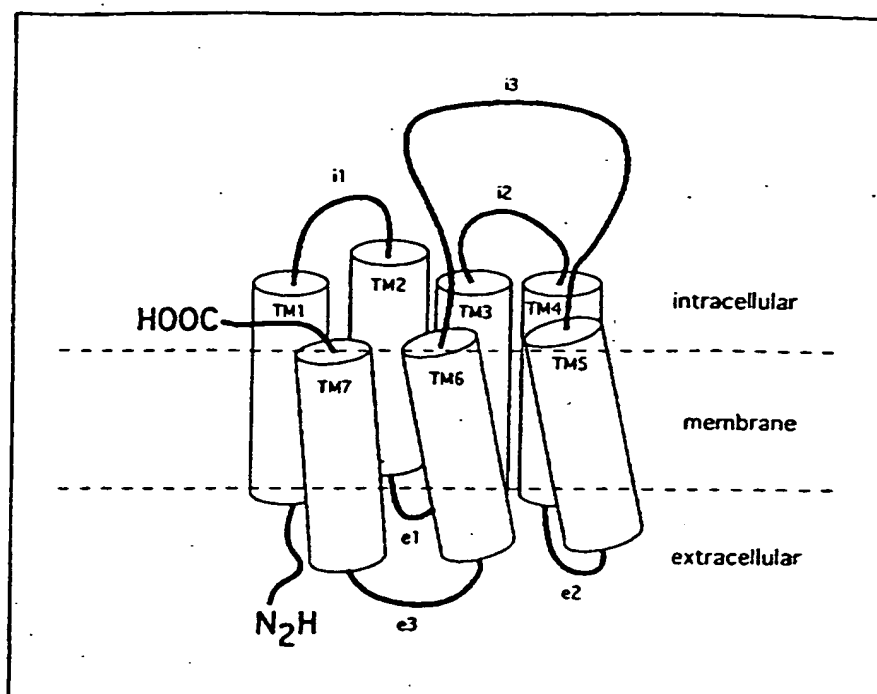


Fig. 1. Model of the structural domains of G-protein-coupled receptors. The transmembrane domains are depicted as cylinders perpendicular to the plane of the plasma membrane. Transmembrane domains 1-7 (TM1-TM7) are proposed to traverse the membrane in an alpha-helical fashion and be connected by alternating extracellular (e1-e3) and intracellular (i1-i3) loops. The amino- (NH₂) and carboxyl- (COOH) terminal regions of G-protein-coupled receptors are situated at the extracellular and intracellular sides of the plasma membrane, respectively.

having seven transmembrane α -helices connected by hydrophilic loops, with the transmembrane domains being arranged in bundles perpendicular to the lipid bilayer. In addition, both bacteriorhodopsin and the GPCR rhodopsin contain the light-absorbing molecule 11-*cis*-retinal. A conserved Lys residue found in the same relative position on transmembrane domain 7 (TM7) in bacteriorhodopsin as in rhodopsin serves as the covalent attachment point for the chromophore. Although bacteriorhodopsin does not belong to the family of GPCRs, the structural similarities between these two classes of proteins are noteworthy.

Based on primary sequence analysis, members of the GPCR gene family can be categorized into distinct subfamilies (Figs. 2 and 3). These include receptors that bind the biogenic amines (e.g., epinephrine, dopamine, acetylcholine), glycoprotein hormones (e.g., thyrotropin, follicle-stimulating hormone, lutropin/choriogonadotropin) and neurokinins (e.g., substance P, sub-

stance K, neuromedin K). The recent cloning of the calcitonin, parathyroid hormone and secretin receptors represents the delineation of yet another subfamily of GPCRs. These receptors are more closely related to one another (up to 42% sequence identity) than to any of the other seven transmembrane-spanning GPCRs (less than 12%).¹⁰⁻¹² In many instances, a receptor within a subfamily can be further divided into subtypes, each encoded by a separate gene. For example, muscarinic acetylcholine receptors (mAChRs) comprise at least five distinct subtypes (m1, m2, m3, m4, and m5).¹³ Similarly, discrete molecular subtypes of the dopamine receptor have been described (D₁, D₂, D₃, D₄, D₅).¹⁴

During the past five years, considerable insights have been gained into the structure-function relationship of GPCRs through the construction of mutant receptor genes.^{1,15} Inferences about receptor structure and function have been deduced from the phenotypes of the mutant proteins. *In vitro* mutage

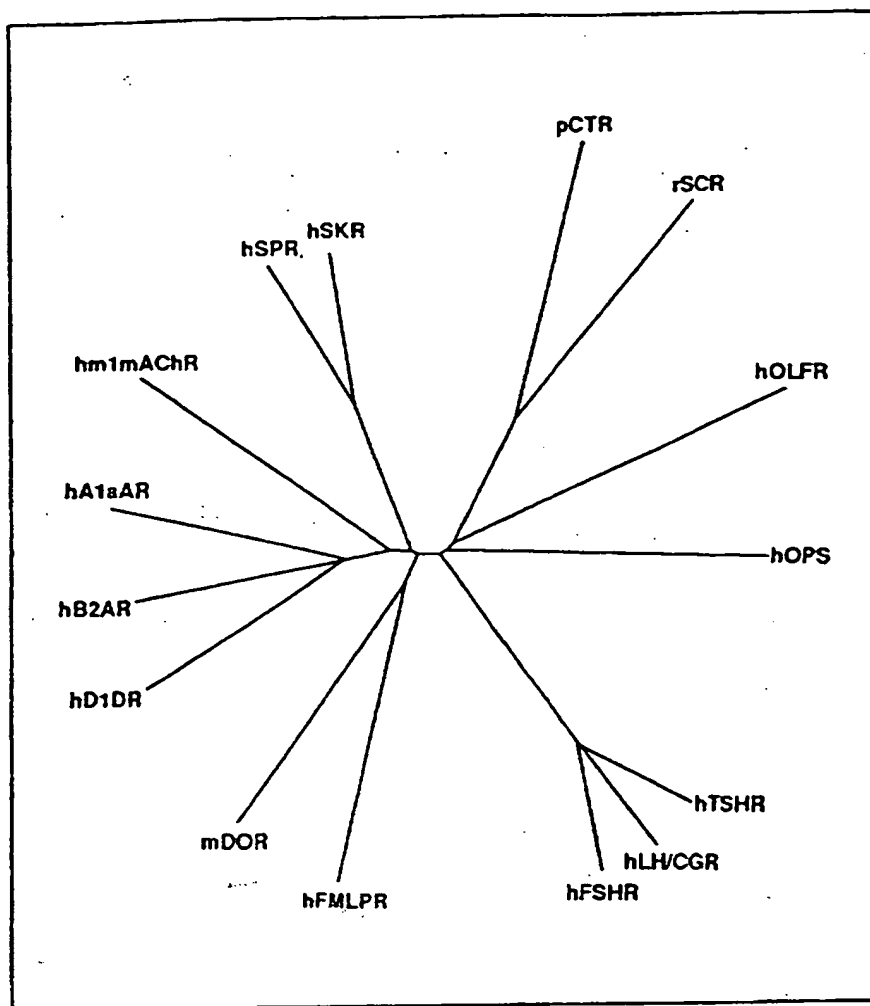


Fig. 2. Relative homology of G-protein-coupled receptors. Sequences were aligned using CLUSTAL⁶⁰ and refinements to the alignment were made manually. The dendrogram was created using the DeSoete Tree Fit⁶¹ and TreeTool (Mike Maciukenas, University of Illinois, unpublished). Only the aligned transmembrane regions were used in the distance calculations. The lengths of the lines are proportional to the percent difference between any two given sequences. All programs were run using the Genetic Data Environment (Steve Smith, Harvard University, unpublished). The considered receptors are as follows: hm1mAChR, human m1 muscarinic⁶²; hA1aAR, human α_{1A} -adrenergic⁶³; hB2AR, human β_2 -adrenergic⁶⁴; hD1DR, human D₁ dopamine⁶⁵; mDOR, mouse delta-opiate⁶⁶; hFMLPR, human *N*-formyl peptide⁶⁷; hSKR, human substance K⁶⁸; hSPR, human substance P⁶⁹; hFSHR, human follicle stimulating hormone⁷⁰; hLH/CHR, human lutropin/choriogonadotropin⁷¹; hTSHR, human thyrotropin⁷²; hOLFR, human olfactory⁷³; hOPS, human rhodopsin⁵; pCTR, porcine calcitonin¹⁰; hSKR, human secretin.¹²

nesis of GPCRs has been used to 1) identify the amino acids critical for ligand binding; 2) determine the domains on the receptor responsible for interacting with G-proteins; and 3) analyze the molecular basis of receptor desensitization. By using molecular modeling techniques in conjunction with information gained by mutational analysis, a better understanding of the roles played by various regions of the receptor protein will provide the rationale for future drug design.¹⁶⁻¹⁸

Amino-terminal domain and extracellular loops

An interesting aspect concerning a number of GPCRs is the apparent lack of an amino-terminal signal peptide sequence. The signal peptide has been demonstrated to be essential for the proper function of integral and secreted proteins,¹⁹ suggesting that an internal signal sequence must exist on those GPCRs lacking an amino-terminal one. In contrast, for GPCRs containing a large amino-terminal domain (more than 300 amino acids), such as the meta-

botropic glutamate receptor (mGluR) and glycoprotein hormone receptors, the presence of a signal sequence has been noted on the amino-terminus.²⁰⁻²² Indeed, the existence of an amino-terminal signal has been confirmed experimentally where the first 26 amino acids deduced from the cDNA sequence of the lutropin/choriogonadotropin receptor (LH/CG-R) are absent on the amino acid sequence derived from purified LH/CG-Rs.²³

Within the amino-terminal domain of all GPCRs are two or more consensus sequences (Asn-X-Ser/Thr) for N-linked glycosylation. For biogenic amine receptors, it is apparent that N-linked glycosylation is not crucial in ligand (agonist and antagonist) binding or receptor/G-protein coupling. For example, treatment of purified β -adrenergic receptors (β AR) with endoglycosidases to remove carbohydrate moieties has no apparent effect on the ligand binding and coupling properties of the reconstituted receptor.^{24,25} Inhibitors of N-linked glycosylation (e.g., tunicamycin) are equally impotent in affecting ligand binding to newly synthesized receptors.²⁶ Similar results are seen with the expression of mutant β ARs and mAChRs.^{27,28} It is likely that glycosylation is essential for the subcellular distribution of some, but not necessarily all, GPCRs. In the case of β ARs, mutant receptors lacking consensus glycosylation sites do not traffic correctly to the cell surface.²⁷ Whether the trafficking defect is due to a decrease in the translocation of receptors from internal stores to the cell surface or an increase in the rate of cell surface receptor internalization remains to be resolved.

GPCRs whose endogenous ligands are biogenic amines lack significant amino-terminal domains (less than 50 amino acids). Early studies focused attention on this region as a potential candidate for the ligand binding domain. The role of the extracellular domains in ligand binding is best exemplified by the β AR, a prototypical biogenic amine receptor. When solubilized β AR is treated with proteolytic enzymes, the

	TM1	TM2	TM3	TM4
hβ1AR	AGHGLMALIVLLIVAGNVLVIVA	FIMSLASADLVHGLLVVPPGATIV	SFFCELMTSDVVLCTVAGIETLCVIALDRY	ARGLVCTVMAISALVSFLPLIM
hβ2AR	VGHGIVMSLIVLAIVFGNVLVITA	FITSLACADLVHGLLVVPPGAAHI	NWCFEFTWSDVVLCTVAGIETLCVIAVDRT	ARVVIIMHVIIVSGLTSTF.LPIQ
ham α1bAR	ISVGLVLCAGFILFAIVGNVLVITS	FIVNLAIAEDLLSTLVPFSATLE	RIFCDTHAAVDVLCCTASILSLCAISIDRY	AAILALLSVWVLSTVISIG.PLL
hα2aAR	LTLVCLAGLLHLLTVFQNVLTIA	FLVSLASADLVHGLLVVPPGAAHI	KTWCEIYALDVLCTSSIVHLCASIDRY	KAI.IITCVVISAVISFPLIS
hα1DR	ILTACTFLSLLILSTLIGNVLVCAA	FVISLAVSLLVAVLVNPKVAE	GSGFNTHAVFDHCTSTASILNLCVSDRT	AFILISVAVTSLVLSIFIPVL
hSHT1aR	VITSLLGLTILFCAVLGNACVAA	LIGSLAVTLNHSVVLVPHAAVYQ	QVTCDEPLALDVLCTTSILHLCALDRY	PRALISLTWLTIGFLISIP.PML
hα1bAChR	AFIGITTCGLSLATVTVGNLVLIS	FLSLACADLIIGTFSNGLTYTL	TLACDLHGLDVLVSNASVWNLISFDRY	AALHGLAWLVSVFLHAPAILF
hα2bAChR	VFTVLVAGSLSVTLITGILVHVS	FLSLACADLIIGTFSNGLTYTL	PVVCDEPLALDVLVSNASVWNLISFDRY	AGNHIAAAWVLSFILHAPAILF
mDOR	IAITALYSAVCAVGLGLNVLVHFG	YIFNLALADALATSTLPFSQAKYL	ELLCKAVLSIDYTNHFTSIFTLTHNSVDRT	AKLINICINWLASGVGVPIHVM
hFHLPR	IITTLVFAVTVFLVGLVGLVTVW	SYLNLAVADFCFTSTLPFFNVRKA	WFLCKFLTFIVDINLFGSVFLTALALDR	AKKVIIGPMVHALLLTLFVIR
hSPR	VLMAAATTVIVVTVVGVVVMVVI	FLVNLAFEAASHAAFMVNVFTYA	LFTYCKFINFFPIAAVFASITSYHTAVAFDRY	TKVVICVWVLALLAFPGCTY
hSKR	ALHAPAYLALVLVAVTGHAVTWE	FIVNLAADLCHAAFNAAFNFYA	RAFCTFQNLFPITAHFVSIYSHTAIAADRY	TKAVIAGIWLVALASPGCTY
mTRHR	VVTLLVWVVCGLGIVGSHVVLV	VLVSLAVADLVHGLLVVPPGAAHI	WGLCLTTLVQLYGLINASCSCTAFTIERY	AKKIIIFVWAFSTIYCHLEFFL
hLH/CGR	DFLRVLMILNITLHGMVTVFL	LHCNLSFADFCNGLVILLIASVDS	GSGGTAGFTTVFASLSEVYTLTVITLERN	AILNLGGLVSSLIAMHPLVG
hFSHR	NILRVLMFISILATGHTIVLVI	LHCNLAFAADLCIGVILLIASVDI	CAOCDAGFTTVFASLSEVYTLTVITLERN	AASVWVWGTFAFAALFPFPG
hTSHR	KFLRIVVWVLSLLALLGSHVFLI	LHCNLAFAADFCNGLVILLIASVDL	GPQNTAGFTTVFASLSEVYTLTVITLERN	ACAINVGGWVCCFLALLPLVG
rmGluR1	IIAIAFSCGLILVTLFVTLIFVLY	YIILAGIFLGYVC.PFTLIAPPT	YLQRLVLGSSAMCYSALVTKTNRIARILA	OVIILASILISVQLTVLVTLIIM
rmGluR2	VGPVTTACGLALATLFLVGVFVRH	YTLGCVFLCT.CHITVFIAPST	TLRRGLGLTAFSVCSALLTKTNRIARIFG	QVAICLALISQGLITVAAMLV
	TM5	TM6	TM7	
hβ1AR	AYAIASSVVFVFLCIMAFTVYL	TLGIINGVFTLCWLPFFLANV	DRLVFFVFWLGYANBAFNPFIYC	
hβ2AR	AYAIASSVVFVFLVIMVFTVY	TLGIINGVFTLCWLPFFVNIH	KEVYILLNIGTVNSGFFPLIYC	
ham α1bAR	FYALFSSLCFTPLAVILVNYC	TLGIIVGHPILCWLPTLALPL	DAVKFVFWLGYFNSCLMFIITP	
hα2aAR	WYVSSICICFPAECILHILVTV	TLGIIVGHPILCWLPTLALPL	DAVKFVFWLGYFNSCLMFIITP	
hα1DR	YTAISSVVISFTIPVAINVITTY	TLGIVMGVPCWLPFFILNCI	SMTDFVFWGFWANSLMFIITA	
hSHT1aR	GYTIYSTGAFITPILLMLVLYG	TLGIINGVFTLCWLPFFVNIH	TLGAIINWLGYSNLLMFIITA	
hα1bAChR	IITFGTAAAFPLVTVNCTLYM	TLGAILLAPILTWTPIHNVV	ETLMELGYLGVNBTIMPCTA	
hα2bAChR	AVTFGTAAAFPLVITHTVLYM	TILAILLAPITWAPVNMVLI	NTVMTIGVWLYINSTIMPCTA	
mDOR	VTKICVFLFAFVVPILITVCTG	MVLVVGAFFVWMAPIHIFTV	VAALHCLTALGYANSLMFIITA	
hFHLPR	VRGITRIFIGFSAPHSIVAVSYG	VLSFVAAPFLCWSPTQVVALI	GIAVDVTSALAFFNSCLMFIITP	
hSPR	VYHICVTVLIXELPLVIGVAYT	SHIVVVCIPAICWLPFFIFLL	QOYVLAHMLANSSTNYMFIITC	
hSKR	LYHLVVIALIXELPLVAVFAYS	THVLVVLTPAICWLPFFIYFIL	QOYVLAHMLANSSTNYMFIITC	
mTRHR	PITLMDFGVGVNPHILATVLYG	HLAVVVILFALLMHPITLVV	NWFLFCRICIYLNBAIMPVITN	
hLH/CGR	YILTILILNVVAFFIICACYIKI	KMAILIFTDFCHAPISFFAIS	TNSKILLVLFYPINSCAMPFLYA	
hFSHR	YVSHLLVWVLAFFVIGCCYIHI	RMAHLIFTDFCHAPISFFAIS	SKAKILLVLFYPINSCAMPFLYA	
hTSHR	YIVFLVTNIVAFVIVCCCHVXI	RMAVLIIFTDFCHAPISFFALS	SNSKILLVLFYPINSCAMPFLYA	
rmGluR1	LCGVAPVGYNGLLIMSCITYAFK	AFTNYTTCIIMLAFPIYFGSN	CFAVLSVTVAGCNETPKNYII	
rmGluR2	ASNLGSLAYWLLIALCTLYAFK	GFTNYTTCIIMLAFPIFYVTS	CVSVLSGSSVVLGCLFAPKLHII	

Fig. 3. Aligned amino acid sequences of the seven transmembrane domains (TM1–7) and adjacent residues of G-protein-coupled receptors. Bold residues represent highly conserved amino acids. Shaded residues represent conserved residues within a subfamily of receptors. The considered receptors are as follows: hβ1AR, human β1-adrenergic⁷⁴; ham α1bAR, hamster α1b-adrenergic⁷⁵; hα2aAR, human α2a-adrenergic⁴⁴; hSHT1aR, human 5-HT_{1A}⁷⁶; mTRHR, mouse thyrotropin releasing hormone⁷⁷; rmGluR1, rat metabotropic glutamate receptor 1²²; and rmGluR2, rat metabotropic glutamate receptor 2.²² References for remaining sequence data can be found in Figure 2.

resulting hydrophobic core retains its capacity to bind the antagonist [¹²⁵I]-iodocyanopindolol (ICYP).²⁹ Furthermore, the tryptic core is still able to activate the G-protein G_s in response to agonists, which suggests that the hydrophilic extracellular regions of the receptor are not crucial for receptor-ligand interactions. Utilization of genetic techniques has further delineated the role of the extracellular domains on biogenic amine receptors in ligand binding. Deletion mutagenesis of the β₂AR revealed that, for the most part, the amino- and carboxyl-terminal domains and e1, e2 and e3 do not contribute to the binding of ICYP and the agonist isoproterenol.^{30,31} In contrast, re-

moval of any of the transmembrane domains practically abolishes ligand binding. It is apparent from these studies that the binding domain of at least one subfamily of GPCRs (biogenic amine receptors) does not involve the extracellular hydrophilic regions, but actually resides in the transmembrane domains. The same is likely true for the receptors that bind small peptide hormones, but confirmation awaits future experiments. For the glycoprotein hormone receptors, the large amino-terminus (more than 300 amino acids) contains 14 imperfect Leu-rich repeat domains.^{20,21,23} It is thought that the large glycoprotein hormones (28–38

kDa) bind to this repeat structure before interacting secondarily with the membrane-spanning regions. Through the construction of chimeric receptors between members of this receptor subfamily, the extracellular amino-terminal domain has been established as the ligand binding site.^{32,33} In fact, the extracellular domain of the LH/CH-R (minus the remainder of protein) can be expressed in transfected cells that consequently bind choriogonadotropin with high affinity.³⁴

A structural feature shared by all GPCRs is the presence of a conserved Cys residue on e1 and another on e2.

These residues have been implicated in the formation of a disulfide bond, since replacement of either one of these residues at position 106 (Cys106) or 184 (Cys184) with Val in the β_2 AR produces a mutant receptor with altered agonist binding properties.³⁰ Similarly, mutation of Cys98 or Cys178 in the m1mAChR, and Cys110 or Cys187 in rhodopsin, completely abolishes ligand binding.^{35,36} Peptide sequencing of the m1mAChR has confirmed the involvement of these conserved Cys residues in disulfide bond formation.³⁷ From these studies, it is believed that the disulfide linkage does not participate directly in ligand binding *per se*, but rather serves a physical role by maintaining the tertiary structure of GPCRs.

Conserved amino acids in the transmembrane domains

Comparison of the deduced amino acid sequences of members of the GPCR gene family has led to the identification of conserved residues located in several transmembrane domains (Fig. 3). Some residues appear to be globally conserved in the majority of GPCRs despite major structural differences in the endogenous ligands that bind to this family of receptors (e.g., catecholamines, peptides, glycoprotein hormones). One hypothesis is that these highly conserved residues play a common functional or structural role, for example, in the process of receptor activation. Among the conserved residues found throughout the GPCRs are the Gly-Asn pair in TM1, a Leu-Ala-X-X-Asp-Leu motif in TM2, an almost canonical motif of Asp-Arg-Tyr at the TM3/i2 junction, an invariable Trp residue in TM4, and a Pro residue flanked by aromatic amino acids in TM5, TM6 and TM7. Interestingly, the secretin/parathyroid hormone/calcitonin receptor subfamily and the mGluR subfamily are practically devoid of these conserved amino acids. In contrast to the globally conserved residues, other amino acids (found only in a subfamily of GPCRs) are postulated to be involved in receptor class-specific functions, such as the binding of biogenic amine ligands. These include the conserved Asp residue in TM3 and a Ser-X-X-Ser mo-

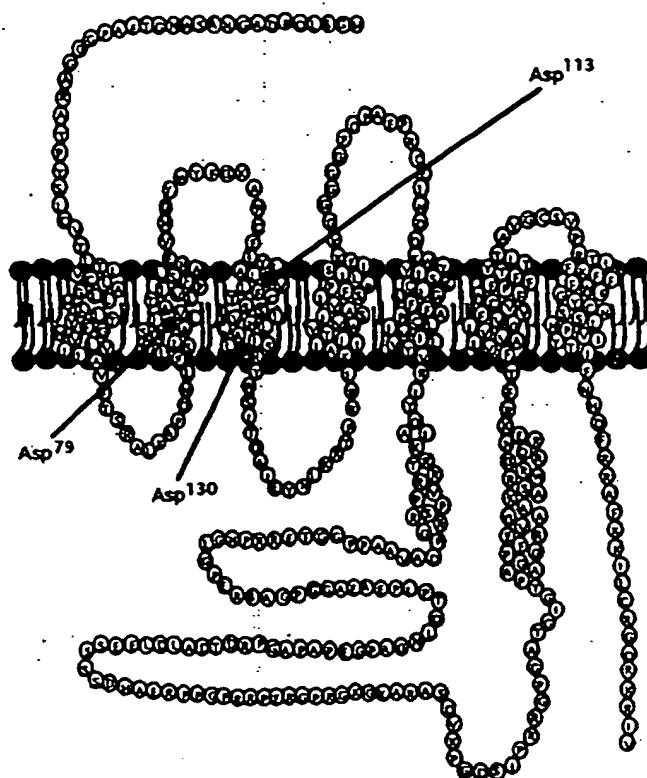
tif in TM5 of the biogenic amine receptors (Fig. 3).

As mentioned above, one of the general features common to a majority of GPCRs is the presence of a pair of Asp residues, one located in the Leu-Ala-X-X-Asp-Leu motif of TM2 and the other situated in the Asp-Arg-Tyr motif at the junction of TM3 and i2 (Fig. 4). The importance of these two Asp residues in receptor function has been well documented for the β_2 AR, m1mAChR, α_2 A-adrenergic receptor (α_2 AR) and dopamine D₁ receptor.³⁸⁻⁴⁰ Expression of the human α_2 AR gene in Chinese hamster ovary cells, cells that normally lack endogenous adrenergic receptors, leads to a pertussis toxin-sensitive inhibition of adenylyl cyclase activity following epinephrine exposure.⁴¹ In pertussis toxin-pretreated cells, however, agonist-mediated activation of α_2 AR leads to an increase in cAMP levels.⁴¹ Substitution of Asp79 with asparagine (Asn) in TM2 of the α_2 AR produces a mutant receptor displaying high-affinity agonist binding and relatively normal antagonist binding properties.³⁸ However, the ability of adrenergic agonists to attenuate adenylyl cyclase activity, as well as enhance cAMP levels in pertussis toxin-pretreated cells, is abolished. Consistent with the inability of agonist to activate mutant [Asn79] α_2 ARs was the observed lack of guanine nucleotide-sensitive high-affinity agonist binding. Asp130 at the TM3/i2 junction of the α_2 AR also appears to influence receptor/G-protein coupling. Mutation of this residue to Asn eliminates high-affinity, guanine nucleotide-sensitive agonist binding. Moreover, agonist-mediated inhibition of adenylyl cyclase activity is markedly attenuated, while elevation of cAMP levels is abolished in pertussis toxin-treated cells.

Similar Asp-to-Asn mutations in the corresponding positions of the β_2 AR and m1mAChR either significantly attenuate or completely eliminate the ability of the mutant receptors to activate adenylyl cyclase and phospholipase C activities, respectively.⁴²⁻⁴⁴ On

the other hand, the effects of these mutations on ligand binding were nominal. Whereas muscarinic agonist and antagonist binding are relatively unaffected in the [Asn71]m1mAChR and [Asn122]m1mAChR mutants, adrenergic agonist affinity is decreased in [Asn79] β_2 AR and slightly increased in [Asn130] β_2 ARs.⁴²⁻⁴⁴ Taken together, these studies suggest that the conserved Asp residues in TM2 and at the TM3/i2 junction are crucial for agonist-induced receptor activation or receptor conformational changes. It has previously been speculated that these invariant negatively charged residues may bind cations and serve as a "charge relay system" during receptor activation by agonists.⁴⁵ It is plausible that the movement of these ions is key to receptor conformational changes following agonist binding. In fact, Asp79 is known to be involved in sodium-dependent allosteric regulation of α_2 AR.⁴⁶ Interestingly, the mutant [Asn79] α_2 AR was found to couple to inhibition of adenylyl cyclase and calcium currents but not to potassium channel activation in AtT20 mouse pituitary tumor cells, suggesting that α_2 ARs undergo different conformations to couple to different G-proteins.⁴⁷

There exists in many G-protein-coupled receptors an Asp residue situated near the extracellular side of TM3 (Fig. 4). Replacement of Asp113 with Asn in the α_2 AR abolishes yohimbine binding and markedly decreases agonist stimulation of the mutant receptor.³⁸ Mutation of the corresponding Asp residue in both β_2 ARs and m1mAChRs likewise affects ligand binding.⁴²⁻⁴⁴ It is unlikely that mutation of this residue alters normal receptor processing and insertion into the lipid bilayer, since [Asn113] β_2 AR can be detected by immunoblotting in membrane preparations.⁴⁸ These findings are consistent with the hypothesis that this Asp residue that is conserved among all biogenic amine receptors, including the α AR, β AR, mAChR, dopamine receptor and serotonin receptor, is involved in an



		79	113	130
α -Adrenergic	Hamster α_1	FIVNIAIA	LLSFTVL	PFSATLE
	Human α_2	FLVSLASAD	LLVATLV	IPFSLANE
β -Adrenergic	Human β_1	FIMSLASAD	VMGLLV	VPGATIV
	Human β_2	FITSLACAD	VMGLAV	VPGAAHI
	Rat β_2	FITSLACAD	VMGLAV	VPGASHI
Muscarinic	Rat m1	FLSLACAD	LIIGTFS	MNLYTTYL
	Rat m2	FLSLACAD	LIIGVFS	MNLYTTYL
	Rat m3	FLSLACAD	LIIGVISH	NLTFTYI
	Rat m4	FLSLGCA	LIIGAFS	MNLYTTYI
	Rat m5	YLLSLACAD	LIIGIFS	MNLYTTYI
Dopamine	Rat D2	LIVSLAVAD	LLVATLV	MPPWVYLE
5-Hydroxytryptamine	Human 5-HT _{1A}	LIGSLAVTD	MVSVLV	PMAALYQ
	Rat 5-HT _{1C}	FLMSLAIA	DLVGLLV	HPSLSLAI
	Rat 5-HT ₂	FLMSLAIA	DLGLLV	MPVSNLTI
		Transmembrane II	Loop	Transmembrane III

Fig. 4. Conservation of aspartate residues in TM2 and TM3 among members of the biogenic amine receptor subfamily. The numbering and location of the conserved aspartate residues are depicted using a model of the human α_2 -adrenergic receptor. References for the sequence data can be found in reference 1. (From: Wang, C.-D., Buck, M.A. and Fraser, C.M. Mol Pharmacol 1991, 40: 168-79; reproduced with permission.)

electrostatic interaction with the cationic amine moiety of their respective ligands.

Ser residues in TM5 are conserved as a pair (Ser-X-X-Ser motif) among biogenic amine receptors that bind catecholamines but not in those receptors

whose endogenous ligand lacks a catechol moiety (e.g., acetylcholine) (Fig. 5). Structure-function analysis of the β_2 AR has implicated the hydroxyl side-chain of Ser204 and Ser207 in hydrogen bond formation with the *meta*- and *para*-hydroxyl groups of catecholamines.⁴⁹ Substitution of either Ser resi-

due with alanine (Ala) attenuates the activity of catecholamine agonists at the mutant receptors. The effects of these mutations on agonist activity can be mimicked by the interaction of *meta*- and *para*-hydroxyl-substituted analogs with the wild-type receptor. Hence, at the [Ala204] β_2 AR mutant, isoprotere-

Human β_1 AR	AYAIASSVVSFYVPLCIMA FVYL
Human β_2 AR	AYAIASSIVSFLVPLVIMV FVYS
Human α_{2A} AR	WYLSSCIGSF FAPCLIMGLVYA
Human α_2 (C-4)	WYLSSCIGSF FAPCLIMGLVYL
Hamster α_1 AR	FYALFSSLGSFYIPLAVILVMYC
Rat D ₂ DR	AFVVYSSIVSFYVPFIVTLLVYI
Dros octop	GYVIYSSLGSFFIPIA IMTIVYI
Rat 5HT-1A	GYTIYSTFGAFYIPLLLMLVLYG
Rat 5HT-1C	NFVLIGSFVAFFIPLTIMVITYF
Human m1	IITFGTAMAAFYLPVTVMCTLYN
Human m2	AVTEGTAIAAFYLPVIIMTVLYW
Human opsin	SFVIYMFVWHFIIPLIVIFFCYG

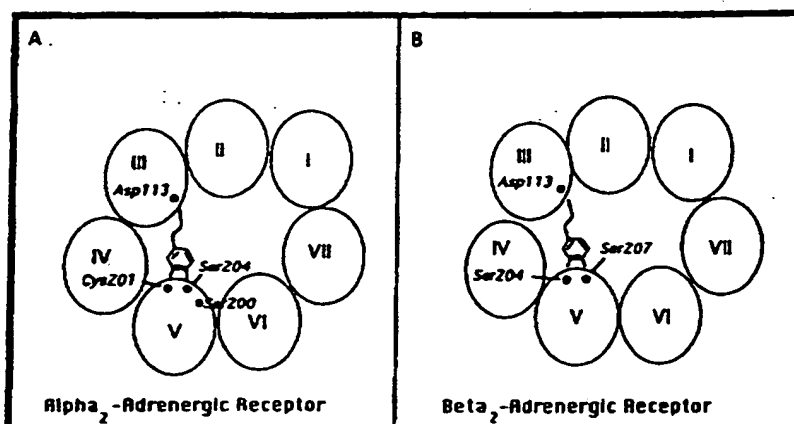


Fig. 5. Conservation of serine residues in TM5 among certain members of the biogenic amine receptor subfamily. Top: Alignment of the deduced amino acid sequences from TM5 of selected G-protein-coupled receptors. Amino acid sequences were aligned to maximize homologies within this region. *, conserved serine residues in a Ser-X-X-Ser motif. References for the sequence data can be found in reference 1. Bottom: Model comparing the ligand binding site of the α_{2A} -adrenergic (A) and β_2 -adrenergic (B) receptors. The view of the receptors is from the extracellular face of the plasma membrane. The seven α -helices are numbered I-VII. Locations of the conserved aspartate (Asp) and serine (Ser) residues implicated in ligand binding are indicated. Ligand binding model for the β_2 -adrenergic receptor has been adapted from reference 48 or 51. (From: Wang, C.-D., Buck, M.A. and Fraser, C.M. Mol Pharmacol 1991, 40: 168-79; reproduced with permission.)

nol and its *meta*-substituted analog display only partial agonist activity, whereas the *para*-substituted analog exhibits no intrinsic agonist activity. Conversely, isoproterenol and its *para*-substituted analog show partial agonist activity at the [Ala207] β_2 AR mutant, but the *meta*-substituted analog is devoid of activity.

In a somewhat analogous manner to that of the [Ala204] β_2 AR mutant, when Ser204 is substituted with Ala in the α_{2A} AR, epinephrine and phenylephrine (*meta*-substituted) elicit 100% maximal agonist activity at the mutant receptor, whereas synephrine (*para*-substituted)

displays only partial agonist activity.³⁸ Based on these findings, it was postulated that Ser204 in the α_{2A} AR functions in a manner similar to that of Ser207 in the β_2 AR by participating in hydrogen bond formation with the *para*-hydroxyl group from the catechol ring structure of catecholamine agonists. There exists a second Ser residue four amino acids upstream from Ser204 in the α_{2A} AR (Fig. 5). Mutation of this residue at position 200 of the α_{2A} AR produces a mutant receptor that is fully activated by epinephrine, phenylephrine and synephrine.³⁸ Thus, Ser200 appears not to directly participate in the ligand binding process. This finding is

not totally unexpected, since Ser204 and Ser207 in the β_2 AR are located three positions apart in TM5, which is presumed to form an α -helix, compared with a distance of four residues apart for Ser200 and Ser204 in the α_{2A} AR. Since one turn of an α -helix encompasses 3.6 amino acids, the hydroxyl group of Ser200 in the α_{2A} AR would assume a different orientation in the helix compared with Ser204 in the β_2 AR. Thus, it is possible that the *meta*-hydroxyl group of catecholamine agonists interacts with the sulfhydryl side-chain of Cys at position 201 of the α_{2A} AR, which is located in the same relative position in TM5 as Ser204 of the β_2 AR (Fig. 5).

C-terminal domain and the intracellular loops

It has been widely presumed that the cytoplasmic loops of GPCRs form an interface between the receptor and G-protein. Several lines of evidence, involving both biochemical and genetic approaches, now lend support for this hypothesis. Findlay and Pappin⁵⁰ revealed early on that proteolytic digestion of i3 of rhodopsin abolished its interaction with the G-protein transducin, thus implicating this domain as the major constituent involved in the coupling process. This finding has been extended to the biogenic amine receptor subfamily through the use of deletion and site-directed mutagenesis. When a large 33-amino-acid deletion (residues 229-258), corresponding to the middle segment of i3, is performed on the hamster β_2 AR, no detectable affect on the ability of the receptor to stimulate adenylyl cyclase was seen.⁵¹ However, deletion of the amino- (222-229) and carboxyl- (258-270) terminal portions of this loop caused marked reductions in agonist-dependent stimulation of adenylyl cyclase.⁵¹ These two short peptide segments are believed to form amphipathic helices that interact with G_s during the process of receptor activation.

Several mutations made by O'Dowd et al.⁵² indicate that other regions on the β_2 AR protein, besides portions of i3, may contribute to receptor coupling. Deletions in i1 and i2

produced mutant receptors with reduced capacity to couple to G_s and stimulate adenylyl cyclase, thereby presupposing a role for these two loops in receptor-G-protein interactions. Furthermore, mutation of a conserved Cys residue (position 341) in the cytoplasmic tail of the β_2 AR impaired the ability of isoproterenol to stimulate adenylyl cyclase.⁵³ Cys341 undergoes palmitoylation and the fatty acid moiety is proposed to insert itself into the lipid bilayer, thus creating an additional cytoplasmic loop. The amino-terminal segment of this "fourth" intracellular loop is speculated to play a role in the coupling of the β AR to G_s , presumably by maintaining proper orientation of the other G-protein binding domains.⁵³

Data obtained on glycoprotein hormone receptors support the general notion of multiple intracellular regions participating in the coupling process. Site-directed mutagenesis of the thyrotropin receptor provides evidence on the importance of i1 and the carboxyl-terminal portions of both i2 and i3 in signal transduction.⁵⁴ In contrast, deletion of two thirds of the carboxyl-terminal end of the cytoplasmic tail does not functionally impair the thyrotropin receptor.⁵⁴ It is not known with certainty whether the remaining amino-terminal portion of the tail, like in the β_2 AR,⁵³ is important in receptor-G-protein coupling.

Chimeric receptors have been constructed to identify the intracellular regions important for defining selective receptor/G-proteins interactions. Studies with chimeric m1/m2 or m2/m3 mAChRs indicate that i3 is sufficient in determining the selective coupling of these receptor subtypes to their respective effector enzymes.^{55,56} Similar findings have been reported for chimeric α_2/β_2 - and β_2/α_1 -ARs.^{57,58} However, it is likely that multiple cytoplasmic domains are required for G-protein binding specificity. Wong et al.⁵⁹ have shown that substitution of a 12-amino-acid segment (in the amino-terminus of i3) of the β_1 AR into the corresponding position of the m1mAChR is enough to

confer G_s , without disturbing G_p , coupling to the latter receptor. Only upon additional substitution of the corresponding i2 domains was G_p coupling to the m1mAChR abolished.⁵⁹ Hence, these data demonstrate the pivotal, although not exclusive, role of i3 in selective effector coupling.

Concluding remarks

The rapid proliferation and identification of newly cloned GPCRs reveal a much greater diversity within this supergene family than was previously considered at the pharmacological level. As more receptors are cloned, the use of site-directed mutagenesis in conjunction with molecular modeling techniques will help better define the functional domains of these proteins. Ultimately, it is this knowledge which will form the basis for the development of future therapeutics.

Acknowledgments

We are grateful to Dr. Claire M. Fraser for critical review of the manuscript.

References

1. Savarese, T.M. and Fraser, C.M. *In vitro* mutagenesis and the search for structure-function relationships among G protein-coupled receptors. *Biochem J* 1992, 283: 1-19.
2. Gilman, A.G. Transducers of receptor-generated signals. *Ann Rev Biochem* 1987, 56: 615-49.
3. Birnbaumer, L. G proteins in signal transduction. *Annu Rev Pharmacol Toxicol* 1990, 30: 675-705.
4. Ovchinnikov, Y.A. Rhodopsin and bacteriorhodopsin: Structure-function relationships. *FEBS Lett* 1982, 148: 179-91.
5. Nathans, J. and Hogness, D.S. Isolation and nucleotide sequence of the gene encoding human rhodopsin. *Proc Natl Acad Sci USA* 1984, 81: 4851-5.
6. Kerlavage, A.R. G-protein-coupled receptor family. *Current Opin Struct Biol* 1991, 1: 393-401.
7. Kerlavage, A.R., Fraser, C.M. and Venter, J.C. Muscarinic cholinergic receptor structure: Molecular biological support for subtypes. *Trends Pharmacol Sci* 1987, 8: 426-31.
8. Henderson, R. and Unwin, P.N.T. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* 1975, 257: 28-32.
9. Henderson, R., Baldwin, J.M., Ceska, T.A., Zemlin, F., Beckmann, E. and Downing, K.H. Model for the structure of bacteriorhodopsin based on high-resolution electron cryomicroscopy. *J Mol Biol* 1990, 213: 899-929.
10. Lin, H.Y., Harris, T.L., Flannery, M.S., Aruffo, A., Kaji, E.H., Gorn, A., Kolakowski, Jr., L.F., Lodish, H.F. and Goldring, S.R. Expression cloning of an adenylyl cyclase-coupled calcitonin receptor. *Science* 1991, 254: 1022-4.
11. Juppner, H., Abou-Samra, A.-B., Freeman, M., Kong, X.F., Schipani, E., Richards, J., Kolakowski, Jr., L.F., Hock, J., Potts, Jr., J.T., Kronenberg, H.M. and Segre, G.V. A G protein-linked receptor for parathyroid hormone and parathyroid hormone-related peptide. *Science* 1991, 254: 1024-6.
12. Ishihara, T., Nakamura, S., Kaziro, Y., Takahashi, T., Takahashi, K. and Nagata, S. Molecular cloning and expression of a cDNA encoding the secretin receptor. *EMBO J* 1991, 10: 1635-41.
13. Bonner, T.I. The molecular basis of muscarinic receptor diversity. *Trends Neurosci* 1989, 12: 148-51.
14. Sibley, D.R. and Monsama, Jr., F.J. Molecular biology of dopamine receptors. *Trends Pharmacol Sci* 1992, 13: 61-9.
15. Lee, N.H., Pellegrino, S.P. and Fraser, C.M. Site-directed mutagenesis of alpha-adrenergic receptors. *Neuroprotocols* 1993, in press.
16. Hibert, M.F., Trumpp-Kallmeyer, S., Bruinvels, A. and Hotlack, J. Three dimensional models of neurotransmitter G-binding protein-coupled receptors. *Mol Pharmacol* 1991, 40: 8-15.
17. Findlay, J. and Eliopoulos, E. Three-dimensional modelling of G protein-linked receptors. *Trends Pharmacol Sci* 1990, 11: 492-9.
18. Trumpp-Kallmeyer, S., Hoflack, J., Bruinvels, A. and Hilbert, M. Modeling of G-protein-coupled receptors: Applications to dopamine, adrenaline, serotonin, acetylcholine, and mammalian opsin receptors. *J Med Chem* 1992, 35: 3448-62.
19. von Heijne, G. The signal peptide. *J Membrane Biol* 1990, 115: 195-201.
20. Parmentier, M., Libert, F., Maenhaut, C., Lefort, A., Gerald, C., Perret, J., Van Sande, J., Dumont, J.E. and Vassart, G. Molecular cloning of the thyrotropin receptor. *Science* 1989, 246: 1620-2.
21. Sprengel, R., Braun, T., Nikolics, K., Segaloff, D.L. and Seeburg, P.H. The testicular receptor for follicle stimulating hormone: Structure and functional expression of cloned cDNA. *Mol Endocrinol* 1990, 4: 525-30.
22. Tanabe, Y., Masu, M., Ishii, T., Shigemoto, R. and Nakanishi, S. A family of metabotropic glutamate receptors. *Neuron* 1992, 8: 169-79.
23. McParland, K.C., Sprengel, R., Phillips, H.S., Kohler, M., Rosembly, N., Nikolics, K., Segaloff, D.L. and Seeburg, P.H. Lutropin-choriogonadotropin receptor: An unusual member of the G protein-coupled receptor family. *Science* 1989, 245: 494-9.
24. Stiles, G.L., Benovic, J.L., Caron, M.G. and Lefkowitz, R.J. Mammalian beta-adrenergic re-

- ceptors. Distinct glycoprotein populations containing high mannose or complex type carbohydrate chains: *J Biol Chem* 1984, 259: 8655-63.
25. Bezovic, J.L., Staniszewski, C., Cerione, R.A., Codina, J., Lefkowitz, R.J. and Caron, M.G. The mammalian beta-adrenergic receptor: Structural and functional characterization of the carbohydrate moiety. *J Recept Res* 1987, 7: 257-81.
26. Doss, R.C., Kramamarcy, N.R., Harden, T.K. and Perkins, J. Effects of tunicamycin on the expression of beta-adrenergic receptors in human astrocytoma cells during growth and recovery from agonist-induced down-regulation. *Mol Pharmacol* 1985, 27: 507-16.
27. Rands, E., Candelore, M.R., Cheung, A.H., Hill, W.S., Strader, C.D. and Dixon, R.A.F. Mutational analysis of beta-adrenergic receptor glycosylation. *J Biol Chem* 1990, 265: 10759-64.
28. van Koppen, C.J. and Nathanson, N.M. Site-directed mutagenesis of the m2 muscarinic acetylcholine receptor. Analysis of the role of N-glycosylation in receptor expression and function. *J Biol Chem* 1990, 265: 20887-92.
29. Rubenstein, R.C., Wong, S.K.-F. and Ross, E.M. The hydrophobic tryptic core of the beta-adrenergic receptor retains G_s-regulatory activity in response to agonists and thiols. *J Biol Chem* 1987, 262: 16655-62.
30. Dixon, R.A.F., Sigal, I.S., Candelore, M.R., Register, R.B., Scattergood, W., Rands, E. and Strader, C.D. Structural features required for ligand binding to the beta-adrenergic receptor. *EMBO J* 1987, 6: 3269-75.
31. Dixon, R.A.F., Sigal, I.S., Rands, E., Register, R.B., Candelore, M.R., Blake, A.D. and Strader, C.D. Ligand binding to the beta-adrenergic receptor involves its rhodopsin-like core. *Nature* 1987, 326: 73-7.
32. Nagayama, Y., Wadsworth, H.L., Chazenbalk, G.D., Russo, D., Seto, P. and Rapoport, B. Thyrotropin-leuteinizing hormone/chorionic gonadotropin receptor extracellular domain chimeras as probes for thyrotropin receptor function. *Proc Nat Acad Sci USA* 1991, 88: 902-5.
33. Braun, T., Schofield, P.R. and Sprengel, R. Amino-terminal leucine-rich repeats in gonadotropin receptors determine hormone selectivity. *EMBO J* 1991, 10: 1885-90.
34. Xie, Y.-B., Wang, H. and Segaloff, D.L. Extracellular domain of lutropin/choriogonadotropin receptor expressed in transfected cells bind choriogonadotropin with high affinity. *J Biol Chem* 1990, 265: 21411-4.
35. Kamik, S.S. and Khorona, H.G. Assembly of functional rhodopsin requires a disulfide bond between cysteine residues 110 and 187. *J Biol Chem* 1990, 265: 17520-4.
36. Savarese, T.M., Wang, C.-D. and Fraser, C.M. Site-directed mutagenesis of the rat m1 muscarinic acetylcholine receptor. Role of conserved cysteines in receptor function. *J Biol Chem* 1992, 267: 11439-48.
37. Kurtenbach, E., Curtis, C.A.M., Pedder, E.K., Aitken, A., Harris, A.C.M. and Hulme, E.C. Muscarinic acetylcholine receptors. Peptide sequencing identifies residues involved in antagonist binding and disulfide bond formation. *J Biol Chem* 1990, 265: 13702-8.
38. Wang, C.-D., Buck, M.A. and Fraser, C.M. Site-directed mutagenesis of alpha_{2A}-adrenergic receptors: Identification of amino acids involved in ligand binding and receptor activation by agonists. *Mol Pharmacol* 1991, 40: 168-79.
39. Neve, K.M., Cox, B.A., Henningsen, R.A., Spanoyannis, A. and Neve, R.L. Pivotal role for aspartate-80 in the regulation of dopamine D2 receptor affinity for drugs and inhibition of adenylyl cyclase. *Mol Pharmacol* 1991, 39: 733-9.
40. Lee, N.H., Hu, J. and El-Fakahany, E.E. Modulation by certain conserved aspartate residues of the allosteric interaction of gallamine at the m1 muscarinic receptor. *J Pharmacol Exp Ther* 1992, 262: 312-6.
41. Fraser, C.M., Arakawa, S., McCambie, W.R. and Venter, J.C. Cloning, sequence analysis, and permanent expression of a human alpha₂-adrenergic receptor in Chinese hamster ovary cells. Evidence for independent pathways of receptor coupling to adenylyl cyclase attenuation and activation. *J Biol Chem* 1989, 264: 11754-61.
42. Chung, F.-Z., Wang, C.-D., Potter, P.C., Venter, J.C. and Fraser, C.M. Site directed mutagenesis and continuous expression of human beta-adrenergic receptors. Identification of a conserved aspartate residue involved in agonist binding and receptor activation. *J Biol Chem* 1988, 263: 4052-5.
43. Fraser, C.M., Chung, F.-Z., Wang, C.-D. and Venter, J.C. Site-directed mutagenesis of human beta-adrenergic receptors: Substitution of aspartic acid-130 by asparagine produces a receptor with high-affinity agonist binding that is uncoupled from adenylyl cyclase. *Proc Nat Acad Sci USA* 1988, 85: 5478-82.
44. Fraser, C.M., Wang, C.-D., Robinson, D.A., Gocayne, J.D. and Venter, J.C. Site-directed mutagenesis of m1 muscarinic acetylcholine receptors: Conserved aspartic acids play important roles in receptor function. *Mol Pharmacol* 1989, 36: 840-7.
45. Venter, J.C., Fraser, C.M., Kertavage, A.R. and Buck, M.A. Molecular biology of adrenergic and muscarinic cholinergic receptors: A perspective. *Biochem Pharmacol* 1989, 38: 1197-208.
46. Horstman, D.A., Brandon, S., Wilson, A.L., Guyer, C.A., Cragoe, Jr., E.J. and Limbird, L.E. An aspartate conserved among G-protein receptors confers allosteric regulation of alpha₂-adrenergic receptors by sodium. *J Biol Chem* 1990, 265: 21590-5.
47. Surprenant, A., Horstman, D.A., Akbarali, H. and Limbird, L.E. A point mutation of the alpha₂-adrenoceptor that blocks coupling to potassium but not calcium currents. *Science* 1992, 257: 977-80.
48. Strader, C.D., Sigal, I.S., Register, R.B., Candelore, M.R., Rands, E. and Dixon, R.A.F. Identification of residues required for ligand binding to the beta-adrenergic receptor. *Proc Nat Acad Sci USA* 1987, 84: 4384-8.
49. Strader, C.D., Candelore, M.R., Hill, W.S., Sigal, I.S. and Dixon, R.A.F. Identification of two serine residues involved in agonist activation of the beta-adrenergic receptor. *J Biol Chem* 1989, 264: 13572-8.
50. Findlay, J.B.C. and Pappin, D.J.C. The opsin family of proteins. *Biochem J* 1986, 238: 625-42.
51. Strader, C.D., Dixon, R.A.F., Cheung, A.H., Candelore, M.R., Blake, A.D. and Sigal, I.S. Mutations that uncouple the beta-adrenergic receptor from G_s and increase agonist affinity. *J Biol Chem* 1987, 262: 16439-43.
52. O'Dowd, B.F., Hnatowich, M., Regan, J.W., Leader, W.M., Caron, M.G. and Lefkowitz, R.J. Site-directed mutagenesis of the cytoplasmic domains of the human beta₂-adrenergic receptor. Localization of regions involved in G protein-receptor coupling. *J Biol Chem* 1988, 263: 15985-92.
53. O'Dowd, B.F., Hnatowich, M., Caron, M.G., Lefkowitz, R.J. and Bouvier, M. Palmitoylation of the human beta₂-adrenergic receptor. Mutation of cys341 in the carboxyl tail leads to an uncoupled nonpalmitoylated form of the receptor. *J Biol Chem* 1989, 264: 7564-9.
54. Chazenbalk, G.D., Nagayama, Y., Russo, D., Wadsworth, H.L. and Rapoport, B. Functional analysis of the cytoplasmic domains of the human thyrotropin receptor by site-directed mutagenesis. *J Biol Chem* 1990, 265: 20970-5.
55. Kubo, T., Bujo, H., Akiba, I., Nakai, J., Mishina, M. and Numa, S. Location of a region of the muscarinic acetylcholine receptor involved in selective effector coupling. *FEBS Lett* 1988, 241: 119-25.
56. Wess, J., Bonner, T.L., Dorje, F. and Brann, M.R. Delineation of muscarinic receptor domains conferring selectivity of coupling to guanine nucleotide-binding proteins and second messengers. *Mol Pharmacol* 1990, 38: 517-23.
57. Kobilka, B.K., Kobilka, T.S., Daniel, K., Regan, J.W., Caron, M.G. and Lefkowitz, R.J. Chimeric alpha₂-beta₂-adrenergic receptors: Delineation of domains involved in effector coupling and ligand binding specificity. *Science* 1988, 240: 1310-6.
58. Cotecchia, S., Exum, S., Caron, M.G. and Lefkowitz, R.J. Regions of the alpha₁-adrenergic receptor involved in coupling to phosphatidylinositol hydrolysis and enhanced sensitivity of biological function. *Proc Nat Acad Sci USA* 1990, 87: 2896-900.
59. Wong, S.K.-F., Parker, E.M. and Ross, R.M. Chimeric muscarinic cholinergic: beta-Adrenergic receptors that activate G_s in response to muscarinic agonists. *J Biol Chem* 1990, 265: 6219-24.
60. Higgins, D.G. and Sharp, P.M. CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. *Gene* 1988, 73: 237-44.

61. De Soete, G. A least squares algorithm for fitting additive trees to proximity data. *Psychometrika* 1983, 48: 621-6.
62. Peralta, E.G., Winslow, J.W., Peterson, G.L., Smith, D.H., Ashkenazi, A., Ramachandran, J., Schimerlick, M.I. and Capon, D.J. Primary structure and biochemical properties of an M2 muscarinic receptor. *Science* 1987, 236: 600-5.
63. Bruno, J.F., Whittaker, J., Song, J. and Berelowitz, M. Molecular cloning and sequencing of a cDNA encoding a human α_1A adrenergic receptor. *Biochem Biophys Res Commun* 1991, 179: 1485-90.
64. Chung, F.-Z., Lentes, K.-U., Gocayne, J., Fitzgerald, M., Robinson, D., Kerlavage, A.R., Fraser, C.M. and Venter, J.C. Cloning and sequence analysis of the human brain beta-adrenergic receptor. Evolutionary relationship to rodent and avian beta-receptors and porcine muscarinic receptors. *FEBS Lett* 1987, 211: 200-6.
65. Zhou, Q.Y., Grandy, D.K., Thambi, L., Kushner, J.A., Van Tol, H.H. and Cone, R. Cloning and expression of human and rat D1 dopamine receptor. *Nature* 1990, 347: 76-80.
66. Kieffer, B.L., Befort, K., Gaveriaux-Ruff, C. and Hirth, C.G. The delta-opioid receptor: Isolation of a cDNA by expression cloning and pharmacological characterization. *Proc Nat Acad Sci USA* 1992, 89: 12048-52.
67. Boulay, F., Tardif, M., Brouchon, L. and Vignais, P. Synthesis and use of a novel N-formyl peptide derivative to isolate a human N-formyl peptide receptor cDNA. *Biochem Biophys Res Commun* 1990, 168: 1103-9.
68. Gerard, N.P., Eddy, Jr., R.L., Shows, T.B. and Gerard, C. The human neurokinin A (substance K) receptor. Molecular cloning of the gene, chromosomal localization, and isolation of the cDNA from tracheal and gastric tissues. *J Biol Chem* 1990, 265: 20455-62.
69. Takeda, Y., Takeda, J., Sachais, B.S. and Krause, J.E. Molecular cloning, structural characterization and functional expression of the human substance P receptor. *Biochem Biophys Res Commun* 1991, 179: 1232-40.
70. Minegishi, T., Nakamura, K., Takakura, Y., Ibuki, Y. and Igarashi, M. Cloning and sequencing of human FSH receptor cDNA. *Biochem Biophys Res Commun* 1991, 175: 1125-30.
71. Minegishi, T., Nakamura, K., Takakura, Y., Miyamoto, K., Hasegawa, Y., Ibuki, Y. and Igarashi, M. Cloning and sequencing of the human LH/hCG receptor cDNA. *Biochem Biophys Res Commun* 1990, 172: 1049-54.
72. Nagayama, Y., Kaufman, K.D., Seto, P. and Rapoport, B. Molecular cloning, sequence and functional expression of the cDNA for the human thyrotropin receptor. *Biochem Biophys Res Commun* 1989, 165: 1184-90.
73. Buck, L. and Axel, R. A novel mungene family may encode odorant receptors: A molecular basis for odor recognition. *Cell* 1991, 65: 175-87.
74. Frielle, T., Collins, S., Daniel, K.W., Caron, M.G., Lefkowitz, R.J. and Kobilka, B.K. Cloning of the cDNA for the human beta₁-adrenergic receptor. *Proc Nat Acad Sci USA* 1987, 84: 7920-4.
75. Cotecchia, S., Schwinn, D.A., Randall, R.R. and Lefkowitz, R.J. Molecular cloning and expression of the cDNA for the hamster alpha_{1A}-adrenergic receptor. *Proc Nat Acad Sci USA* 1988, 85: 7159-63.
76. Kobilka, B.K., Frielle, T., Collins, S., Yang-Feng, T.L., Kobilka, T.S., Francke, U., Lefkowitz, R.J. and Caron, M.G. An intronless gene encoding a potential member of the family of receptors coupled to guanine nucleotide regulatory proteins. *Nature* 1987, 329: 75-9.
77. Straub, R.E., Frech, G.C., Joho, R.H. and Gershengorn, M.C. Expression cloning of a cDNA encoding the mouse pituitary thyrotropin-releasing hormone receptor. *Proc Nat Acad Sci USA* 1990, 87: 9514-8.

Norman H. Lee, Ph.D., and Anthony R. Kerlavage, Ph.D., are associated with the Dept. of Molecular and Cellular Biology and the Computational Genomics Facility, respectively, of The Institute for Genomic Research, 932 Clopper Rd., Gaithersburg, Maryland 20878, U.S.A.

CHIRON REPORTS PROMISING PRECLINICAL AND CLINICAL RESULTS

In its first quarter report for 1993, Chiron Corp. reported promising preclinical and clinical results for several products.

Interferon beta. Interferon beta-1b was recommended for approval by an FDA advisory panel for the treatment of relapsing/remitting multiple sclerosis. In the area of vaccines, the company reported that results of phase II trial of The Biochemicals Company's (Chiron's joint venture with Eisai) HSV-2 vaccine showed a 20% reduction in the frequency of genital herpes outbreaks. Phase III trials are under way on an

improved HSV-2 vaccine. In a phase I study, vaccinated individuals showed an immune response higher than previous vaccine. If results are confirmed, BioCine will proceed with phase III studies in both immunocompetent and immunodeficient patients.

The National Institutes of Health have awarded Chiron a \$1.5 million contract to develop a vaccine against hepatitis B virus. The vaccine is being tested in high-risk groups, including 10- to 15-year-old children with hepatitis B, AIDS, and hemodialysis patients. Preliminary studies also show that the vaccine is effective in preventing hepatitis B virus infection in chimpanzees. Phase II studies are under way on a vaccine against hepatitis C virus.

rus (HCV) vaccine. BioCine is preparing phase I studies, to begin in 1994.

In oncology, a trial has begun to determine the efficacy of a new type of bi-specific monoclonal antibody that targets the c-erbB-2 oncogene protein and that promotes the antitumor effects of killer cells designed to interact with antibodies. At an FDA-sponsored workshop earlier this year, preliminary results were presented from a phase I trial of platelet-derived growth factor (PDGF) induced by Shurin and Shurin's partner, Elicon, in patients with multiple myeloma. The results showed a 50% reduction in tumor size. BioCine is planning a phase II trial of the compound in patients with multiple myeloma.

The Wall Street Journal
Copyright (c) 2001, Dow Jones & Company, Inc.

Wednesday, January 10, 2001

The Cure: With Big Drugs Dying, Merck Didn't
Merge

--

It Found New Ones

Some Inspired Research, Aided By a Bit of Luck,
Saves Company's Independence

The Path to a Novel Painkiller
By Gardiner Harris
Staff Reporter of The Wall Street Journal

For 15 years, Edward Scolnick, head of Merck & Co.'s drug research, knew the company would be facing a crisis about now. For much of that time, he secretly feared that Merck might not survive it as an independent company. "I had some doubts that I didn't share with anybody," Dr. Scolnick says.

Merck's problem, which at times has infected almost every big pharmaceuticals company, was that patents on several of its best-selling drugs would be expiring. Generic knockoffs would then eat deeply into market share and profits on drugs like Vasotec and Prinivil for hypertension, Mevacor for high cholesterol and Pepcid and Prilosec for ulcers.

Ever since investors caught on to this, Wall Street has been insisting that Merck join the merger rush sweeping the pharmaceuticals industry. But its chief executive, Raymond V. Gilmartin, steadfastly refused, insisting that Merck could grow briskly all by itself.

He was right. Today, well into what was supposed to be the crunch, Merck is riding high. It topped all its peers in revenue growth last year, and most of them in earnings growth, analysts say. Its stock surged 26% in 2000 while the broad market was skidding. Instead of facing an acute need to save money, Merck is increasing its research spending by nearly 17% and its sales force by almost a third.

"The safe thing would have been to seek a merger, emphasize generics, stay diversified and cut costs across the board," Mr. Gilmartin says. "We went

against the conventional wisdom at the time, stayed with it and did it."

Mr. Gilmartin, 59 years old, who arrived at Merck after heading a medical-device company, gambled that the pharmaceuticals giant's tradition of creativity and innovation in drug discovery would bail it out. A merger, by contrast, would dilute the power of this science-based culture -- one that has been a model for other drug companies -- and be a distraction for years.

Had he and his lieutenants been wrong, Merck's name might have wound up in the same graveyard as Warner-Lambert, Upjohn, Syntex, Sandoz, Ciba-Geigy, Rhone-Poulenc and Hoechst, all of which had to resort to mergers when their labs couldn't produce enough new drugs to replace old ones with expired patents.

Merck's success demonstrates that in the drug business, as in Hollywood, one big hit can sway the fate of an entire company. And searching for blockbuster drugs is a matter of inspiration, scientific instincts and shrewd management -- assets that are hard to buy in a merger.

In this case, the inspiration came from Peppi Prasit, a Thai-born medicinal chemist for Merck in Montreal. In July 1992, he found himself wandering around an obscure medical conference in that city. Chatting with a colleague, he learned that Merck researchers had developed a lab test to determine if a painkiller was less likely to cause the stomach upset that goes along with most pain and arthritis drugs.

Moments later, Dr. Prasit, now 45, noticed a poster display from some researchers for a Japanese company claiming they had produced just such a nonirritating painkiller, though one that wasn't chemically fit to try in people. Dr. Prasit immediately went back to his Montreal lab, cooked up the mysterious molecule and put it to Merck's new test. When it passed, he set about trying to create a similar drug for humans.

His work caught the eye of Dr. Scolnick, the research chief at Merck's sprawling laboratory northwest of Philadelphia. Dr. Scolnick, 60, is a former molecular biologist for the National Institutes

of Health who joined Merck in 1982 and became its top scientist three years later. A dour and fierce man, he works out of a small office with a view of industrial pipes and a door so narrow he has to slide sideways to get through. His adjoining conference room is decorated only with two bedraggled plastic plants.

The unimpressive surroundings belie the critical nature of his job: Dr. Scolnick monitors hundreds of intriguing scientific leads floating around Merck's labs and decides where the company will make its big bets. Merck's winnowing process has evolved over three decades into committees of scientists who discuss one another's work with brutal frankness. It's a system of peer review modeled on one used at the NIH, and it "allows a really good debate about what we should be doing," says Dr. Scolnick.

The system has helped Merck, in recent years, to bring out new drugs like Fosomax, to slow bone deterioration in osteoporosis; Singulair for asthma; and big-selling medicines for high blood pressure, glaucoma, migraine and AIDS. But Dr. Scolnick says he didn't need a committee to tell him that Dr. Prasit's painkiller project had the potential to be a blockbuster, and a critical bridge out of Merck's patent problem.

The class of painkillers called nonsteroidal anti-inflammatory drugs -- like aspirin and ibuprofen -- hadn't seen a major improvement in years. And thousands of Americans suffered ulcers each year because of the drugs' side effects. Preventing that would clearly be a huge advance.

These drugs attack the inflammation that leads to pain by curbing production of prostaglandins, compounds that marshal the body's defenses. But prostaglandins also are involved in making the lining that protects the gut from digestive acids. The more a painkiller inhibited inflammation, the more it thinned the protective lining, increasing the risk of bleeding ulcers.

Philip Needleman, a pharmacologist at Washington University in St. Louis, had mapped out a potential way around this. It would be a drug that inhibited Cox-2, an enzyme that regulates prostaglandin production in most of the body, but not Cox-1, a similar enzyme involved only in the gut. Dr. Prasit

knew of this research and was determined to develop just such a drug, especially since Merck had a test for it.

But in this quest, Dr. Prasit and Dr. Scolnick feared they were in a race -- and running second. Rumors swirled that Dr. Needleman, who subsequently crossed town to join Monsanto Co., was working on a similar drug for that company.

So Dr. Scolnick ordered researchers in Montreal to pursue Dr. Prasit's work as fast as they could. "I would call up every other day and say, 'Hey, is everybody working on this project?' " recalls Dr. Scolnick with a rare smile. "They would always say, 'Yes!' You don't know if they're telling the truth, but they got the message that it was important."

Dr. Prasit's team synthesized hundreds of compounds, some of which worked great in the test tube but passed through laboratory mice with no clinical effect. Others mysteriously killed the mice. But by October 1994, the team had come up with two compounds that aced the test-tube tests and didn't hurt the mice, even at extremely high doses.

Normally, Dr. Scolnick would have chosen one of these to put through the expensive and risky process of testing in humans. But the project was so important -- and Merck appeared to be in such a high-stakes competition with Monsanto -- that he decided to put both compounds in clinical trials.

It was a good move, because only one of the two ended up working. "One failed and the other didn't, and there was no way you could have looked at the preclinical data and predicted which one would succeed," Dr. Scolnick says. "That's just dumb luck."

Meanwhile, Mr. Gilmartin, arriving in 1994, had other headaches. Growth at the company, so stellar in the late 1980s, had slowed. A health plan proposed by the Clinton administration threatened price controls. Powerful managed-care organizations were demanding deep discounts. "There were people that were questioning, in a managed-care environment, what was going to be the value of breakthrough research," the CEO says. "Merck, in fact, had even moved into the generics business. Everything was being questioned and challenged."

Among his first moves was to squelch the push into generics and sell off specialty-chemicals and agricultural subsidiaries. He also ordered his managers to make peace with managed-care companies.

Merck had been fighting their demands for discounts, with the result that its products were increasingly being excluded from the "formulary" lists of large buying groups. "Merck was just in your face. If you tried to set up a meeting with them, they would refuse," says Lynn Detlor, president of American Healthcare Systems' Purchasing Partners LP, a huge group-purchasing organization. In Mr. Gilmartin's first week in 1994, he set up a meeting with that group's executives and promised that Merck would cooperate at every level, say two participants at the meeting. Within 18 months, the managed-care group had increased its purchases of Merck products tenfold on an annualized basis, Mr. Detlor says.

But most important, Mr. Gilmartin decided then to bet the company's future on the productivity of its labs. "Shortly after he came here, he came to one of our research meetings and stayed for dinner," Dr. Scolnick recalls. "As he was leaving, on the way out he said he wanted to talk to me. And he said, 'I want you to know that I have complete confidence in you. Just do your thing and I'm not going to bother you.'"

That meant he had freedom to throw all the resources he wanted into Dr. Prasit's project in Montreal.

In January 1995, Merck handed a batch of a potential new pain drug to Donald R. Mehlich, an oral surgeon in Austin, Texas, who tests such medicines for manufacturers. He recruits students from the University of Texas, yanks out their wisdom teeth, gives them a pill and puts them into a dorm attached to his clinic to watch them suffer. "We create a lot of pain in what we do," he says cheerfully.

A test drug's effectiveness is measured largely by how long it takes patients to insist that what they were given isn't working and they need something else. The tests are designed to be "double-blind," with neither doctor nor patient knowing which pill is

which. Even so, Dr. Mehlich sensed that what he was testing for Merck had potential. It was "the first time we've ever had a compound that has worked so well for so long," he says.

Meanwhile, Monsanto, which was working on a similar drug just as Merck had suspected, ran its candidate through similar dental-pain tests. It failed them. However, it and Merck's drug were both good at relieving the longer-term pain of arthritis, without stomach irritation.

To Merck's dismay, Monsanto completed its clinical studies first. Among the reasons was a dosage glitch at Merck. The company figured out only belatedly that, instead of as much as 1,000 milligrams, the proper dose was 12.5 mg. to 25 mg. The pills that resulted were so tiny that Merck was afraid arthritis patients wouldn't be able to pick them up. It enlarged them with edible filler, but that caused another problem -- the filler turned out to slow the drug's absorption. Three months were lost while researchers worked to fix all this.

On the last day of 1998, the Food and Drug Administration gave Monsanto approval to market its nonirritating painkiller, called Celebrex. In February 1999, Monsanto began co-marketing it with Pfizer Inc. -- and it quickly became the most successful drug launch in U.S. history. Merck still didn't even have marketing approval.

Normally, a head start like that makes the first drug dominant and very hard to catch. Yet the way Merck handled its later launch would soon put its drug, called Vioxx, hot on Celebrex's trail. One reason: an expanded role for marketers within Merck.

For decades, Merck's marketers hadn't been allowed anywhere near scientific- planning meetings. Mr. Gilmartin's predecessor, P. Roy Vagelos, started to change this, persuading scientists to accept marketers in their midst by promising that they wouldn't speak. Then, speaking was allowed but not encouraged. Under Mr. Gilmartin, however, the marketers have become deeply involved in many of the scientists' development decisions, though they still have no involvement in early-stage research issues.

Mr. Gilmartin created teams of marketing,

manufacturing and research people that now plan far ahead. "We deconstructed every task to see where we could cut out steps," says Wendy Dixon, a marketing vice president who oversaw the Vioxx launch. "We carved four or five weeks off the normal product-launch process."

While the team made thousands of bottles and boxes in advance, they couldn't do that with the pills' instruction flier; the FDA doesn't bless it until the day of approval. With the rival drug already a hit, Merck's challenge would be to get the approved copy from the FDA to print shops across the U.S. and Puerto Rico, print the fliers by the thousands, insert them with the pills and get the bottles to pharmacies in just days. Planes were placed on standby in case printing plates needed to be rushed to print shops elsewhere.

Insiders knew that May 20, a Friday, was likely to be the day of FDA approval. Cheryl Ramsey-Weldon, the company's top formatter of instruction fliers, waited all day on tenterhooks. When her shift ended she went home and waited some more. "I was sitting by the phone," she says. "I called [her supervisor] three times to see how close we were."

She finally got the call at 10 p.m. Ready for bed with her contacts out, her hair up and her pajamas on, Ms. Ramsey-Weldon jumped into her car as she was and raced the 3 1/2 miles to the plant. Four hours later, she had formatted the document and passed it along to a pair of proofreaders. At 2:30 a.m., she went home for a few hours' sleep. By 6 a.m. she was back for more.

Merck's presses ran for days without stop. Then the fliers were folded and inserted. The bottles reached distribution centers on Monday afternoon. Vioxx was stocked in 40,000 pharmacies within 11 days of approval, a remarkable feat.

Within three months of its launch, the Merck drug gained nearly a third of the brand-new market for "Cox-2 inhibitors," according to research firm IMS Health, and within a year it had nearly half. In Europe, Vioxx is dominant, having beaten Celebrex to market in most countries despite filing later. Helping Vioxx in the heated two-way competition: It acts more quickly than Celebrex and is more selective for the Cox-2 enzyme, according to independent studies.

Both drugs have flaws, though. And now Merck is locked in another Cox-2 contest, racing Pharmacia Inc. -- which took over Monsanto -- to bring out second-generation, improved versions of the hot-selling drugs.

These days, Mr. Gilmartin has an uncharacteristic swagger. "We're going to another level at a time when most worried that Merck wouldn't even compete," he says. But Dr. Scolnick gives plenty of credit to the way things broke Merck's way during Vioxx's development. "If those first two compounds had failed [in human trials] and we had had by chance to rely on the fifth or sixth one" years later, he says, "we would be a very different company."

Prescription for Success

Merck is gradually losing its exclusive rights to these drugs . . .

Drug (condition)	1999 sales U.S. (in millions)	1999 sales World-wide (in millions)	Expiration*
Vasotec (Hypertension)	\$975	\$2,300	August 2000
Mevacor (Cholesterol)	\$480	\$600	December 2001
Pepcid (Ulcers)	\$820	\$910	April 2001
Prinivil (Hypertension)	\$715	\$815	June 2002

But the company has drugs in the pipeline

Launches expected in 2001

-- Cancidas: intravenous anti-fungal drug; application submitted to FDA July 2000

-- Invanz: intravenous antibiotic; application submitted to FDA November 2000

-- Eterocoxib: Super Vioxx for arthritis; application to FDA expected

1/10/01 Wall St. J. A1 2001 WL-WSJ 2850599

early 2001

*End of exclusivity

NOTE: In November, the patent will also expire on Prilosec, an Astra-Zeneca drug for heartburn and gastro-esophageal reflux disease, from which Merck receives considerable revenue.

---- INDEX REFERENCES ----

COMPANY (TICKER): Merck & Co. Inc. (MRK)

NEWS SUBJECT: Newspapers' Section Fronts; Marketing; Front-Page Stories; Page-One Story; Corporate Profiles; Management Issues; People Profile; Research & Development; Research and Development; Dow Jones Total Market Index; Wall Street Journal; Corporate and Industrial News; English language content; Content Types; Health; Political and General News; Health (FRT MRK PAG NPAG PRO C41 NPEO RND C23 WEI WSJ CCAT ENGL NCAT GHEA GCAT HLT)

MARKET SECTOR: Consumer Non-Cyclical (NCY)

INDUSTRY: Drug Manufacturers (DRG)

PRODUCT: Pharmaceuticals; Wall Street Journal Graphics (DPH PIC)

REGION: New Jersey; United States - New Jersey; North America; North America (Regional Focus); United States; United States; Eastern U.S.; North American Countries; Pacific Rim Countries (NJ USNJ NME NAM US USA USE NAMZ PACRMZ)

LAYOUT CODES: Page One Umbrella; Right Leader (PGO RGT)

Word Count: 2637

1/10/01 WSJ A1

END OF DOCUMENT

**This Page is Inserted by IFW Indexing and Scanning
Operations and is not part of the Official Record**

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images include but are not limited to the items checked:

☒ **BLACK BORDERS**

☐ **IMAGE CUT OFF AT TOP, BOTTOM OR SIDES**

☒ **FADED TEXT OR DRAWING**

☒ **BLURRED OR ILLEGIBLE TEXT OR DRAWING**

☐ **SKEWED/SLANTED IMAGES**

☐ **COLOR OR BLACK AND WHITE PHOTOGRAPHS**

☐ **GRAY SCALE DOCUMENTS**

☒ **LINES OR MARKS ON ORIGINAL DOCUMENT**

☐ **REFERENCE(S) OR EXHIBIT(S) SUBMITTED ARE POOR QUALITY**

☐ **OTHER: _____**

IMAGES ARE BEST AVAILABLE COPY.

As rescanning these documents will not correct the image problems checked, please do not report these problems to the IFW Image Problem Mailbox.